

Shallow water equations: Split-form, entropy stable, well-balanced, and positivity preserving numerical methods

Hendrik Ranocha

September 26, 2016

Entropy stable semidiscretisations of the shallow water equations are developed, based on summation-by-parts (SBP) operators and using split forms of the equations. The resulting two-parameter family of entropy conservative schemes for general SBP bases, especially using Gauß nodes, is adapted to varying bottom topography in a well-balanced way, i.e. preserving the lake-at-rest steady state. Moreover, positivity preservation is ensured using the framework of Zhang and Shu (*Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and recent developments*, 2011. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, The Royal Society, vol 467, pp. 2752–2766) and finite volume subcells, adapted to nodal SBP bases with diagonal mass matrix. Numerical tests of the proposed schemes are performed and some conclusions are presented.

1 Introduction

This article is concerned with numerical methods for the shallow water equations in one space dimension based on *summation-by-parts* (SBP) operators, see inter alia the review articles by Fernández, Hicken, and Zingg (2014); Svärd and Nordström (2014) and references cited therein. This setting of SBP operators originates in the *finite difference* (FD) setting, but can also be used in polynomial methods as nodal *discontinuous Galerkin* (DG) (Gassner, 2013) or flux reconstruction / correction procedure via reconstruction (Ranocha, Öffner, and Sonar, 2016).

Entropy stability has long been known as a desirable stability property for conservation laws. Here, the semidiscrete setting of Tadmor (1987, 2003) will be used. Other desirable stability properties for the shallow water equations are the preservation of non-negativity of the water height and the correct handling of steady states, especially the lake-at-rest initial condition, resulting in well-balanced methods.

This article extends the entropy conservative split form of Gassner, Winters, and Kopriva (2016a); Wintermeyer, Winters, Gassner, and Kopriva (2016) to a two-parameter family of well-balanced and entropy conservative splittings, enables the use of bases not including boundary nodes, and considers positivity preservation using the framework of Zhang and Shu (2011) and *finite volume* (FV) subcells. Additionally, some known entropy stable and positivity preserving numerical fluxes are compared and implementation details are provided.

Other references for numerical methods for the shallow water equations can be found in the review article of Xing and Shu (2014) and references cited therein.

At first, some analytical properties of the shallow water equations are reviewed in section 2 and the existing split form of Gassner, Winters, and Kopriva (2016a); Wintermeyer, Winters, Gassner, and Kopriva (2016) is described in section 3. Afterwards, a two-parameter family of entropy conservative numerical fluxes for the shallow water equations with constant bottom topography is developed in section 4 and extended to a varying bottom in section 5. The corresponding semidiscretisation using general SBP bases is designed in section 6. The positivity preserving framework of Zhang and Shu (2011) is introduced to this setting in section 7 and numerical fluxes based on the entropy conserving schemes are investigated with respect to positivity preservation in section 8. Additionally, some known fluxes are presented. An extension of the idea to use FV subcells to the setting of nodal SBP bases with diagonal mass matrix is proposed in section 9 and numerical experiments are presented in section 10. Finally, the results are summed up in section 11 and some conclusions and directions of further research are presented.

2 Shallow water equations

The shallow water equations in one space dimension are

$$\underbrace{\partial_t \begin{pmatrix} h \\ hv \end{pmatrix}}_{=u} + \partial_x \underbrace{\begin{pmatrix} hv \\ hv^2 + \frac{1}{2}gh^2 \end{pmatrix}}_{=f(u)} = \underbrace{\begin{pmatrix} 0 \\ -gh\partial_x b \end{pmatrix}}_{=s(u,x)} \quad (1)$$

where h is the water height, v its speed, hv the discharge, b describes the bottom topography, and g is the gravitational constant.

The entropy for a constant bottom topography b is given by the total energy

$$U = \frac{1}{2}hv^2 + \frac{1}{2}gh^2 = \frac{1}{2}\frac{u_2^2}{u_1} + \frac{1}{2}gu_1^2. \quad (2)$$

The associated entropy variables ($b \equiv \text{const}$) are

$$w = U'(u) = \begin{pmatrix} gu_1 - \frac{1}{2}\frac{u_2^2}{u_1} \\ \frac{u_2}{u_1} \end{pmatrix} = \begin{pmatrix} gh - \frac{1}{2}v^2 \\ v \end{pmatrix}. \quad (3)$$

Thus,

$$w'(u) = U''(u) = \begin{pmatrix} g + \frac{u_2^2}{u_1^3} & -\frac{u_2}{u_1^2} \\ -\frac{u_2}{u_1^2} & \frac{1}{u_1} \end{pmatrix} = \begin{pmatrix} g + \frac{v^2}{h} & -\frac{v}{h} \\ -\frac{v}{h} & \frac{1}{h} \end{pmatrix}, \quad (4)$$

and this is positive definite for $h > 0$, i.e. U is strictly convex for positive water height h . Therefore, the conservative variables u and the entropy variables w can be used alternatively. The flux expressed in terms of the entropy variables ($b \equiv \text{const}$) is

$$f(u(w)) = \begin{pmatrix} \frac{w_1 + \frac{1}{2}w_2^2}{g}w_2 \\ \frac{w_1 + \frac{1}{2}w_2^2}{g}w_2^2 + \frac{1}{2}g\left(\frac{w_1 + \frac{1}{2}w_2^2}{g}\right)^2 \end{pmatrix} = \frac{1}{g} \begin{pmatrix} w_1w_2 + \frac{1}{2}w_2^3 \\ \frac{1}{2}w_1^2 + \frac{3}{2}w_1w_2^2 + \frac{5}{8}w_2^4 \end{pmatrix}, \quad (5)$$

and the flux potential ($b \equiv \text{const}$) is given by

$$\psi = \frac{1}{2g}w_1^2w_2 + \frac{1}{2g}w_1w_2^3 + \frac{1}{8g}w_2^5 = \frac{1}{2}gh^2v, \quad (6)$$

fulfilling

$$\psi'(w) = f(u(w)). \quad (7)$$

The entropy flux for constant bottom topography $b \equiv \text{const}$

$$\begin{aligned} F &= w \cdot f(u(w)) - \psi(w) = gh^2v - \frac{1}{2}hv^3 + hv^3 + \frac{1}{2}gh^2v - \frac{1}{2}gh^2v \\ &= \frac{1}{2}hv^3 + gh^2v \end{aligned} \quad (8)$$

fulfils

$$F'(u) = w'(u) \cdot f(u) + w \cdot f'(u) - \psi'(w) \cdot w'(u) = w \cdot f'(u) = U'(u) \cdot f'(u). \quad (9)$$

Therefore, for smooth solutions with $b \equiv \text{const}$,

$$\begin{aligned} \partial_t U(u) + \partial_x F(u) &= U'(u) \partial_t u + F'(u) \partial_x u = -U'(u) \partial_x f(u) + F'(u) \partial_x u \\ &= (-U'(u) f'(u) + F'(u)) \partial_x u = 0, \end{aligned} \quad (10)$$

and the entropy inequality

$$\partial_t U + \partial_x F \leq 0 \quad (11)$$

will be used as an additional admissibility criterion for weak solutions.

For general bottom topography b , the entropy / total energy is

$$U = \frac{1}{2}hv^2 + \frac{1}{2}gh^2 + ghb, \quad (12)$$

with associated entropy variables

$$w = U'(u) = \begin{pmatrix} g(h+b) - \frac{1}{2}v^2 \\ v \end{pmatrix} \quad (13)$$

and entropy flux

$$F = \frac{1}{2}hv^3 + gh^2v + ghbv. \quad (14)$$

Again, smooth solutions satisfy $\partial_t U + \partial_x F = 0$, and the entropy inequality $\partial_t U + \partial_x F \leq 0$ will be used as an additional admissibility criterion for weak solutions.

The flux Jacobian

$$f'(u) = \partial_u \begin{pmatrix} u_2 \\ \frac{u_2^2}{u_1} + \frac{1}{2}gu_1^2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\frac{u_2^2}{u_1^2} + gu_1 & 2\frac{u_2}{u_1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ gh - v^2 & 2v \end{pmatrix} \quad (15)$$

has the eigenvalues and eigenvectors

$$\lambda_{\pm} = v \pm \sqrt{gh}, \quad v_{\pm} = \begin{pmatrix} 1 \\ \lambda_{\pm} \end{pmatrix}, \quad (16)$$

and can thus be diagonalised for $h > 0$. The entropy Jacobian (cf. (4))

$$\partial_w u = \begin{pmatrix} \frac{1}{g} & \frac{v}{g} \\ \frac{v}{g} & h + \frac{v^2}{g} \end{pmatrix} \quad (17)$$

can be expressed by using a scaling of the eigenvectors in the form proposed by Barth (1999, Theorem 4) as

$$\partial_w u = RR^T, \quad R = \frac{1}{\sqrt{2g}} \begin{pmatrix} 1 & 1 \\ v - \sqrt{gh} & v + \sqrt{gh} \end{pmatrix}. \quad (18)$$

3 Existing split-form SBP method

A general SBP SAT semidiscretisation is obtained by a partition of the domain into disjoint elements. On each element, the solution is represented in some basis, mostly nodal bases. These cells are mapped to a standard element for the following computations. There, the symmetric and positive definite mass matrix $\underline{\underline{M}}$ induces a scalar product, approximating the L_2 scalar product. The derivative is represented by the matrix $\underline{\underline{D}}$. Interpolation to the (two point) boundary of the cell (interval) is performed via the restriction operator $\underline{\underline{R}}$ and evaluation of the values at the right boundary minus values at the left boundary is conducted by the boundary matrix $\underline{\underline{B}} = \text{diag}(-1, 1)$. Together, these operators fulfil the summation-by-parts (SBP) property

$$\underline{\underline{M}} \underline{\underline{D}} + \underline{\underline{D}}^T \underline{\underline{M}} = \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}}, \quad (19)$$

mimicking integration by parts on a discrete level

$$\int_{\Omega} u(\partial_x v) + \int_{\Omega} (\partial_x u)v = uv|_{\partial\Omega}. \quad (20)$$

Here, the notation of Ranocha, Öffner, and Sonar (2015, 2016) has been used. Then, similar to strong form discontinuous Galerkin methods, the semidiscretisation can be written as the sum of volume terms, surface terms, and numerical fluxes at the boundaries.

Gassner, Winters, and Kopriva (2016a) proposed as semidiscretisation of the shallow water equations (1) with continuous bottom topography b in the setting of a discontinuous Galerkin spectral element method (DGSEM) using Lobatto-Legendre nodes in each element, that can be generalised to diagonal norm SBP operators with nodal bases including boundary nodes. Wintermeyer, Winters, Gassner, and Kopriva (2016) extended this setting to two space dimensions, curvilinear grids and discontinuous bottom topographies. In one space dimension on a linear grid, this semidiscretisation can be written using the notation of Ranocha, Öffner, and Sonar (2016) as

$$\begin{aligned} \partial_t \underline{h} &= -\underline{\underline{D}} \underline{h} \underline{v} - \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \left(\underline{f}_h^{\text{num}} - \underline{\underline{R}} \underline{h} \underline{v} \right), \\ \partial_t \underline{h} \underline{v} &= -\frac{1}{2} \left(\underline{\underline{D}} \underline{h} \underline{v}^2 + \underline{h} \underline{v} \underline{\underline{D}} \underline{v} + \underline{v} \underline{\underline{D}} \underline{h} \underline{v} \right) - g \underline{h} \underline{\underline{D}} \underline{h} - g \underline{h} \underline{\underline{D}} \underline{b} \\ &\quad - \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \left(\underline{f}_{hv}^{\text{num}} - \underline{\underline{R}} \underline{h} \underline{v}^2 - \frac{1}{2} g \underline{\underline{R}} \underline{h}^2 + \frac{1}{2} g \{ \{ h \} \}_- \llbracket b \rrbracket_{-\underline{e}_0} + \frac{1}{2} g \{ \{ h \} \}_+ \llbracket b \rrbracket_{+\underline{e}_1} \right), \end{aligned} \quad (21)$$

where \underline{e}_k is the k -th unit vector and for cell i

$$\begin{aligned} \{ \{ h \} \}_- &= \frac{h_{i,0} + h_{i-1,p}}{2}, & \{ \{ h \} \}_+ &= \frac{h_{i,p} + h_{i+1,0}}{2}, \\ \llbracket b \rrbracket_- &= b_{i,0} - b_{i-1,p}, & \llbracket b \rrbracket_+ &= b_{i+1,0} - b_{i,p}. \end{aligned} \quad (22)$$

Here, $h_{i,0}, h_{i,p}$ are the values of h at the first and last node $0, p$ in cell i , respectively, as also drawn in Figure 1.

Using

$$\begin{aligned} f_h^{\text{num}} &= \{ \{ h \} \} \{ \{ v \} \}, \\ f_{hv}^{\text{num}} &= \{ \{ h \} \} \{ \{ v \} \}^2 + \frac{1}{2} g \{ \{ h^2 \} \}, \end{aligned} \quad (23)$$

as numerical (surface) flux, where $\{ \{ a \} \} = \frac{a_- + a_+}{2}$, the resulting scheme

1. conserves the mass in general and the discharge for a constant bottom topography,
2. conserves the total energy which is used as entropy,
3. handles the lake-at-rest stationary state correctly,

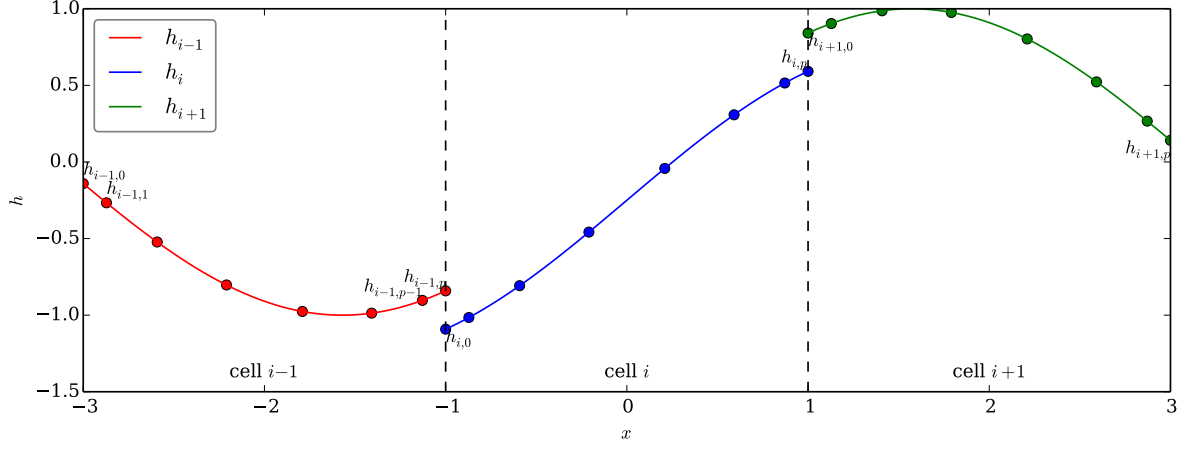


Figure 1: Visualisation of the nodes in the elements $i-1$, i , and $i+1$.

i.e. it is *conservative*, *stable* and *well-balanced*, as proved by Wintermeyer, Winters, Gassner, and Kopriva (2016, Theorem 1). The split form discretisation has been recast into the flux differencing framework of Fisher and Carpenter (2013); Fisher, Carpenter, Nordström, Yamaleev, and Swanson (2013) using the "translations" provided by Gassner, Winters, and Kopriva (2016b, Lemma 1). The resulting volume fluxes are

$$\begin{aligned} f_h^{\text{vol}} &= \{hv\}, \\ f_{hv}^{\text{vol}} &= \{hv\} \{v\} + g \{h\}^2 - \frac{1}{2}g \{h^2\}, \end{aligned} \quad (24)$$

i.e. the volume terms in (21) can be rewritten using the differentiation matrix $\underline{\underline{D}}$ as

$$\begin{aligned} \left[\underline{\underline{D}} hv \right]_i &= \sum_{k=0}^p 2D_{ik} \{hv\}_{ik}, \\ \frac{1}{2} \left[\underline{\underline{D}} hv^2 + hv \underline{\underline{D}} v + v \underline{\underline{D}} hv \right]_i &= \sum_{k=0}^p 2D_{ik} \{hv\}_{ik} \{v\}_{ik}, \\ \left[\underline{\underline{h}} \underline{\underline{D}} h \right]_i &= \sum_{k=0}^p 2D_{ik} \left(\{h\}_{ik}^2 - \frac{1}{2} \{h^2\}_{ik} \right), \end{aligned} \quad (25)$$

where

$$\{a\}_{ik} = \frac{a_i + a_k}{2}. \quad (26)$$

The split form in (21) corresponds to the entropy conservative fluxes f^{vol} . f^{num} corresponds to a splitting, too, but exchanging f^{num} for f^{vol} does not yield a well-balanced method. Similarly, exchanging f^{vol} for f^{num} as surface flux does not work properly. The rest of this paper is dedicated to the investigation of the following questions

1. Are there other entropy conservative fluxes than f^{vol} , f^{vol} and corresponding split forms?
2. Are there other discretisations of $gh\partial_x b$ that can be used to get a well-balanced scheme, respecting the lake-at-rest stationary state?
3. Can the split forms be used for a nodal SBP method without boundary nodes, e.g. for Gauß nodes?
4. Are there entropy conservative / stable and positivity preserving numerical fluxes that can be used to apply the bound preserving framework of Zhang and Shu (2011)? If so, is the resulting method still entropy stable and well-balanced?

4 Entropy conservative fluxes and split forms for vanishing bottom topography $b \equiv 0$

In this section, several numerical fluxes and associated split forms of the shallow water equations (1) with constant bottom topography $b \equiv 0$ will be considered.

The numerical flux $f^{\text{num}} = f^{\text{num}}(u_-, u_+)$ has to be *consistent*, i.e.

$$f^{\text{num}}(u, u) = f(u). \quad (27)$$

In order to be *entropy conservative*, the condition

$$[w] \cdot f^{\text{num}}(u_-, u_+) = [\psi] \quad (28)$$

of Tadmor (1987, 2003) has to be fulfilled in a semidiscrete setting. Similarly, if

$$[w] \cdot f^{\text{num}}(u_-, u_+) \leq [\psi], \quad (29)$$

the numerical flux is *entropy stable*, since it contains more dissipation than an entropy conservative flux. Here, $[a] = a_+ - a_-$.

4.1 Primitive variables h, v

Using a discrete analogue of the product rule

$$\begin{aligned} [ab] &= a_+b_+ - a_-b_- = \frac{a_+ + a_-}{2}(b_+ - b_-) + (a_+ - a_-)\frac{b_+ + b_-}{2} \\ &= \{a\} [b] + [a] \{b\}, \end{aligned} \quad (30)$$

the jump of the flux potential $\psi = \frac{1}{2}gh^2v$ can be written both as

$$[\psi] = \frac{1}{2}g [h^2v] = \frac{g}{2} \left(\{h^2\} [v] + \{v\} [h^2] \right) = \frac{g}{2} \left(\{h^2\} [v] + 2 \{h\} \{v\} [h] \right) \quad (31)$$

and as

$$\begin{aligned} [\psi] &= \frac{1}{2}g [h^2v] \\ &= \frac{g}{2} \left(\{hv\} [h] + \{h\} [hv] \right) = \frac{g}{2} \left(\{hv\} [h] + \{h\}^2 [v] + \{h\} \{v\} [h] \right). \end{aligned} \quad (32)$$

Thus,

$$[\psi] = \frac{1}{2}g [h^2v] = \frac{g}{2}\alpha \left(\{h^2\} [v] + 2 \{h\} \{v\} [h] \right) + \frac{g}{2}(1-\alpha) \left(\{hv\} [h] + \{h\}^2 [v] + \{h\} \{v\} [h] \right), \quad (33)$$

where $\alpha \in \mathbb{R}$ is a real weight of the previous two forms, obtained by the product rule (30). Therefore, using the entropy variables $w = \left(gh - \frac{1}{2}v^2, v \right)^T$ (13), the entropy conservation condition $[w] \cdot f^{\text{num}} = [\psi]$ (28) can be recast as

$$\begin{aligned} &g [h] f_h^{\text{num}} - \frac{1}{2} [v^2] f_h^{\text{num}} + [v] f_{hv}^{\text{num}} \\ &= \frac{g}{2} \alpha \left(\{h^2\} [v] + 2 \{h\} \{v\} [h] \right) + \frac{g}{2} (1-\alpha) \left(\{hv\} [h] + \{h\}^2 [v] + \{h\} \{v\} [h] \right). \end{aligned} \quad (34)$$

Using the product rule (30), $[v^2] = 2 \{v\} [v]$, and this can be restated as

$$\begin{aligned} &g \left(f_h^{\text{num}} - \frac{1+\alpha}{2} \{h\} \{v\} - \frac{1-\alpha}{2} \{hv\} \right) [h] \\ &+ \left(f_{hv}^{\text{num}} - f_h^{\text{num}} \{v\} - g \frac{\alpha}{2} \{h^2\} - \frac{1-\alpha}{2} g \{h\}^2 \right) [v] = 0. \end{aligned} \quad (35)$$

Obviously, a possible choice for f^{num} fulfilling this condition is

$$\begin{aligned} f_{\alpha,h}^{\text{num}} &= \frac{1-\alpha}{2} \{hv\} + \frac{1+\alpha}{2} \{h\} \{v\}, \\ f_{\alpha,hv}^{\text{num}} &= \frac{1-\alpha}{2} \{hv\} \{v\} + \frac{1+\alpha}{2} \{h\} \{v\}^2 + g \frac{\alpha}{2} \{h^2\} + \frac{1-\alpha}{2} g \{h\}^2. \end{aligned} \quad (36)$$

Expanding the terms, this fluxes can be written as

$$\begin{aligned} f_{\alpha,h}^{\text{num}} &= \frac{3-\alpha}{8} (h_+ v_+ + h_- v_-) + \frac{1+\alpha}{8} (h_+ v_- + h_- v_+), \\ f_{\alpha,hv}^{\text{num}} &= \frac{1+\alpha}{8} g (h_+^2 + h_-^2) + \frac{1-\alpha}{4} g h_+ h_- \\ &\quad + \frac{3-\alpha}{16} (h_+ v_+^2 + h_- v_-^2) + \frac{1+\alpha}{16} (h_+ v_-^2 + h_- v_+^2) + \frac{h_+ v_+ v_-}{4} + \frac{h_- v_+ v_-}{4}. \end{aligned} \quad (37)$$

This proves

Lemma 1. *The one-parameter family (36) of numerical fluxes f_α^{num} is a family of consistent and entropy conservative numerical fluxes for the shallow water equations (1) with vanishing bottom topography $b \equiv 0$.*

The numerical surface and volume fluxes f^{vol} (24), f^{num} (23) are members of this family with parameters $\alpha = -1$ and $\alpha = 1$, respectively. Therefore, this one-parameter family of entropy conservative fluxes (36) can also be seen as linear combinations of the two fluxes (23) and (24) with coefficients summing up to one.

Using the translation rules of Gassner, Winters, and Kopriva (2016b, Lemma 1), the fluxes (36) correspond to the split form

$$\begin{aligned} f_{\alpha,h}^{\text{num}} &: \frac{1-\alpha}{2} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} + \frac{1+\alpha}{2} \cdot \frac{1}{2} (\underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} + \underline{\underline{h}} \underline{\underline{D}} \underline{\underline{v}} + \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{h}}) \\ &= \frac{3-\alpha}{4} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} + \frac{1+\alpha}{4} (\underline{\underline{h}} \underline{\underline{D}} \underline{\underline{v}} + \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{h}}) \\ f_{\alpha,hv}^{\text{num}} &: \frac{1-\alpha}{2} \cdot \frac{1}{2} (\underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}}^2 + \underline{\underline{h}} \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{v}} + \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}}) \\ &\quad + \frac{1+\alpha}{2} \cdot \frac{1}{4} (\underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}}^2 + \underline{\underline{h}} \underline{\underline{D}} \underline{\underline{v}}^2 + 2 \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} + \underline{\underline{v}}^2 \underline{\underline{D}} \underline{\underline{h}} + 2 \underline{\underline{h}} \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{v}}) \\ &\quad + g \frac{\alpha}{2} \underline{\underline{D}} \underline{\underline{h}}^2 + \frac{1-\alpha}{2} g \cdot \frac{1}{2} (\underline{\underline{D}} \underline{\underline{h}}^2 + 2 \underline{\underline{h}} \underline{\underline{D}} \underline{\underline{h}}) \\ &= \frac{3-\alpha}{8} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}}^2 + \frac{1}{2} \underline{\underline{h}} \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{v}} + \frac{1}{2} \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} + \frac{1+\alpha}{8} \underline{\underline{h}} \underline{\underline{D}} \underline{\underline{v}}^2 + \frac{1+\alpha}{8} \underline{\underline{v}}^2 \underline{\underline{D}} \underline{\underline{h}} \\ &\quad + \frac{1+\alpha}{4} g \underline{\underline{D}} \underline{\underline{h}}^2 + \frac{1-\alpha}{2} g \underline{\underline{h}} \underline{\underline{D}} \underline{\underline{h}}. \end{aligned} \quad (38)$$

It is also possible to start with a general ansatz of the split form and use conditions for consistency, conservation, and entropy stability (similar to section 6), in order to determine the coefficients. This yields the same one-parameter family of fluxes and corresponding split forms, but is much more tedious.

4.2 Entropy variables w

Similarly to the previous section, the flux potential ψ can also be expressed as a polynomial in the entropy variables w instead of the primitive variables h, v . Therefore, the jump of the flux potential $\psi = \frac{1}{2} g h^2 v = \frac{1}{2g} w_1^2 w_2 + \frac{1}{2g} w_1 w_2^3 + \frac{1}{8g} w_2^5$ (6) can be written as

$$[\psi] = \frac{1}{2g} [w_1^2 w_2] + \frac{1}{2g} [w_1 w_2^3] + \frac{1}{8g} [w_2^5]. \quad (39)$$

Using the product rule (30), the first jump term can be written in two different ways, similar to the previous section. Weighting both variants with weights $a_1 \in \mathbb{R}$ and $1 - a_1$, respectively, results in

$$[w_1^2 w_2] \stackrel{a_1}{=} \{w_1^2\} [w_2] + \{w_2\} [w_1^2] = \{w_1^2\} [w_2] + 2 \{w_1\} \{w_2\} [w_1] \quad (40)$$

$$\begin{aligned}
& \stackrel{1-a_1}{=} \{w_1 w_2\} [w_1] + \{w_1\} [w_1 w_2] \\
& = \{w_1 w_2\} [w_1] + \{w_1\}^2 [w_2] + \{w_1\} \{w_2\} [w_1] \\
& = ((1 + a_1) \{w_1\} \{w_2\} + (1 - a_1) \{w_1 w_2\}) [w_1] \\
& \quad + (a_1 \{w_1^2\} + (1 - a_1) \{w_1\}^2) [w_2],
\end{aligned}$$

for $a_1 \in \mathbb{R}$. Similarly, the second jump term can be expressed in four different ways as

$$\begin{aligned}
& [w_1 w_2^3] \tag{41} \\
& \stackrel{a_2}{=} \{w_2^3\} [w_1] + \{w_1\} [w_2^3] = \{w_2^3\} [w_1] + \{w_1\} \{w_2^2\} [w_2] + 2 \{w_1\} \{w_2\}^2 [w_2] \\
& \stackrel{a_3}{=} \{w_2^2\} [w_1 w_2] + \{w_1 w_2\} [w_2^2] \\
& = \{w_2^2\} \{w_2\} [w_1] + \{w_1\} \{w_2^2\} [w_2] + 2 \{w_1 w_2\} \{w_2\} [w_2] \\
& \stackrel{a_4}{=} \{w_2\} [w_1 w_2^2] + \{w_1 w_2^2\} [w_2] \\
& = \{w_2^2\} \{w_2\} [w_1] + 2 \{w_1\} \{w_2\}^2 [w_2] + \{w_1 w_2^2\} [w_2] \\
& = \{w_2\} [w_1 w_2^2] + \{w_1 w_2^2\} [w_2] \\
& = \{w_2\}^2 [w_1 w_2] + \{w_1 w_2\} \{w_2\} [w_2] + \{w_1 w_2^2\} [w_2] \\
& = \{w_2\}^3 [w_1] + \{w_1\} \{w_2\}^2 [w_2] + \{w_1 w_2\} \{w_2\} [w_2] + \{w_1 w_2^2\} [w_2] \\
& = (a_2 \{w_2^3\} + (a_3 + a_4) \{w_2\} \{w_2^2\} + (1 - a_2 - a_3 - a_4) \{w_2\}^3) [w_1] \\
& \quad + ((a_2 + a_3) \{w_1\} \{w_2^2\} + (1 + a_2 - a_3 + a_4) \{w_1\} \{w_2\}^2 \\
& \quad + (1 - a_2 + a_3 - a_4) \{w_1 w_2\} \{w_2\} + (1 - a_2 - a_3) \{w_1 w_2^2\}) [w_2],
\end{aligned}$$

where $a_2, a_3, a_4 \in \mathbb{R}$. However, expanding the terms in the last expression results in

$$\begin{aligned}
& a_2 \{w_2^3\} + (a_3 + a_4) \{w_2\} \{w_2^2\} - (a_2 + a_3 + a_4) \{w_2\}^3 \\
& = \frac{3a_2 + a_3 + a_4}{8} (w_{2+}^3 - w_{2+}^2 w_{2-} - w_{2+} w_{2-}^2 + w_{2-}^3),
\end{aligned} \tag{42}$$

and

$$\begin{aligned}
& (a_2 + a_3) \{w_1\} \{w_2^2\} + (a_2 - a_3 + a_4) \{w_1\} \{w_2\}^2 \\
& \quad - (a_2 - a_3 + a_4) \{w_1 w_2\} \{w_2\} - (a_2 + a_3) \{w_1 w_2^2\} \\
& = \frac{3a_2 + a_3 + a_4}{8} (-w_{1+} w_{2+}^2 + w_{1+} w_{2-}^2 + w_{1-} w_{2+}^2 - w_{1-} w_{2-}^2).
\end{aligned} \tag{43}$$

Thus, the expression depends only on $3a_2 + a_3 + a_4$ and can be simplified by setting $a_3 = a_4 = 0$ to

$$\begin{aligned}
& [w_1 w_2^3] = (a_2 \{w_2^3\} + (1 - a_2) \{w_2\}^3) [w_1] \\
& \quad + (a_2 \{w_1\} \{w_2^2\} + (1 + a_2) \{w_1\} \{w_2\}^2 \\
& \quad + (1 - a_2) \{w_1 w_2\} \{w_2\} + (1 - a_2) \{w_1 w_2^2\}) [w_2]
\end{aligned} \tag{44}$$

Finally, the last jump term can be expressed as

$$\begin{aligned}
& [w_2^5] \tag{45} \\
& \stackrel{a_5}{=} \{w_2^4\} [w_2] + \{w_2\} [w_2^4] = \{w_2^4\} [w_2] + 4 \{w_2^2\} \{w_2\}^2 [w_2] \\
& \stackrel{a_6}{=} \{w_2^4\} [w_2] + \{w_2\} [w_2^4] = \{w_2^4\} [w_2] + \{w_2\}^2 [w_2^3] + \{w_2\} \{w_2^3\} [w_2] \\
& = \{w_2^4\} [w_2] + 2 \{w_2\}^4 [w_2] + \{w_2\}^2 \{w_2^2\} [w_2] + \{w_2\} \{w_2^3\} [w_2] \\
& = \{w_2^3\} [w_2^2] + \{w_2^2\} [w_2^3] \\
& = 2 \{w_2\} \{w_2^3\} [w_2] + 2 \{w_2\}^2 \{w_2^2\} [w_2] + \{w_2^2\}^2 [w_2]
\end{aligned}$$

$$= \left((a_5 + a_6) \{w_2^4\} + (2 + 2a_5 - a_6) \{w_2\}^2 \{w_2^2\} + 2a_6 \{w_2\}^4 \right. \\ \left. + (2 - 2a_5 - a_6) \{w_2\} \{w_2^3\} + (1 - a_5 - a_6) \{w_2^2\}^2 \right) \llbracket w_2 \rrbracket,$$

where $a_5, a_6 \in \mathbb{R}$. However, these parameters a_5, a_6 are also redundant, since the last expression can be simplified as

$$\begin{aligned} \llbracket w_2^5 \rrbracket &= \left(w_{2+}^4 + w_{2+}^3 w_{2-} + w_{2+}^2 w_{2-}^2 + w_{2+} w_{2-}^3 + w_{2-}^4 \right) \llbracket w_2 \rrbracket \\ &= \left(\{w_2^4\} + 4 \{w_2\}^2 \{w_2^2\} \right) \llbracket w_2 \rrbracket. \end{aligned} \quad (46)$$

Inserting these forms in the condition (28) for an entropy conservative flux,

$$\begin{aligned} f_{a_1, a_2, h}^{\text{num}}(w_-, w_+) &= \frac{1 + a_1}{2g} \{w_1\} \{w_2\} + \frac{1 - a_1}{2g} \{w_1 w_2\} + \frac{a_2}{2g} \{w_2^3\} + \frac{1 - a_2}{2g} \{w_2\}^3, \\ f_{a_1, a_2, hv}^{\text{num}}(w_-, w_+) &= \frac{a_1}{2g} \{w_1^2\} + \frac{1 - a_1}{2g} \{w_1\}^2 + \frac{a_2}{2g} \{w_1\} \{w_2^2\} \\ &\quad + \frac{1 + a_2}{2g} \{w_1\} \{w_2\}^2 + \frac{1 - a_2}{2g} \{w_1 w_2\} \{w_2\} + \frac{1 - a_2}{2g} \{w_1 w_2^2\} \\ &\quad + \frac{1}{8g} \{w_2^4\} + \frac{1}{2g} \{w_2\}^2 \{w_2^2\}, \end{aligned} \quad (47)$$

turns out to be an entropy conservative numerical flux expressed in terms of the entropy variables $w = \left(gh - \frac{1}{2}v^2, v \right)^T$ (3) for vanishing bottom topography $b \equiv 0$. Expanding the terms in entropy and primitive variables, respectively, results after direct but tedious calculations in

$$\begin{aligned} f_{a_1, a_2, h}^{\text{num}} &= \frac{3 - a_1}{8g} (w_{1+} w_{2+} + w_{1-} w_{2-}) + \frac{1 + a_1}{8g} (w_{1+} w_{2-} + w_{1-} w_{2+}) \\ &\quad + \frac{1 + 3a_2}{16g} (w_{2+}^3 + w_{2-}^3) + \frac{3 - 3a_2}{16g} (w_{2+}^2 w_{2-} + w_{2+} w_{2-}^2), \\ f_{a_1, a_2, hv}^{\text{num}} &= \frac{1 + a_1}{8g} (w_{1+}^2 + w_{1-}^2) + \frac{1 - a_1}{4g} w_{1+} w_{1-} + \frac{7 - 3a_2}{16g} (w_{1+} w_{2+}^2 + w_{1-} w_{2-}^2) \\ &\quad + \frac{1 + 3a_2}{16g} (w_{1+} w_{2-}^2 + w_{1-} w_{2+}^2) + \frac{w_{1+} w_{2+} w_{2-}}{4g} + \frac{w_{1-} w_{2+} w_{2-}}{4g} \\ &\quad + \frac{1}{8g} (w_{2+}^4 + w_{2+}^3 w_{2-} + w_{2+}^2 w_{2-}^2 + w_{2+} w_{2-}^3 + w_{2-}^4), \end{aligned} \quad (48)$$

and

$$\begin{aligned} f_{a_1, a_2, h}^{\text{num}} &= \frac{3 - a_1}{8} (h_+ v_+ + h_- v_-) + \frac{1 + a_1}{8} (h_+ v_- + h_- v_+) \\ &\quad + \frac{a_1 + 3a_2 - 2}{16g} (v_+^3 - v_+^2 v_- - v_+ v_-^2 + v_-^3), \\ f_{a_1, a_2, hv}^{\text{num}} &= \frac{1 + a_1}{8} g (h_+^2 + h_-^2) + \frac{1 - a_1}{4} g h_+ h_- - \frac{2a_1 + 3a_2 - 5}{16} (h_+ v_+^2 + h_- v_-^2) \\ &\quad + \frac{2a_1 + 3a_2 - 1}{16} (h_+ v_-^2 + h_- v_+^2) + \frac{h_+ v_+ v_-}{4} + \frac{h_- v_+ v_-}{4} \\ &\quad + \frac{a_1 + 3a_2 - 2}{32g} (v_+^4 - 2v_+^2 v_-^2 + v_-^4). \end{aligned} \quad (49)$$

This proves

Lemma 2. *The two-parameter family (47) of numerical fluxes $f_{a_1, a_2}^{\text{num}}$, expressed also as (48) and (49), is a family of consistent and entropy conservative numerical fluxes for the shallow water equations (1) with vanishing bottom topography $b \equiv 0$.*

Comparing the numerical fluxes obtained by a splitting of primitive variables (37) and entropy variables (49), the flux f_α^{num} (36) is a special case of the flux $f_{a_1, a_2}^{\text{num}}$ (47) with parameters $a_2 = \frac{2-a_1}{3}$ and $a_1 = \alpha$.

Using entropy variables, the numerical flux (48) does not seem to be qualitatively different from the one-parameter family. However, expressing the terms in primitive variables (49) reveals that the one-parameter family (37) may be more relevant, since no higher order terms are introduced.

The crucial ingredient to obtain the entropy conservative fluxes in Lemmas 1 and 2 has been the expression of both the entropy variables w and the flux potential ψ as polynomials in the same variables, either primitive variables h, v or entropy variables w . Therefore, it may be conjectured, that if such a condition is complied with, there are entropy conservative fluxes expressed in terms of these variables, corresponding to a split form as described inter alia by Gassner, Winters, and Kopriva (2016b) and in section 6.

Furthermore, the general entropy conservative flux of Tadmor (1987, Equation (4.6a)), obtained by integration in phase space, can be recovered by the family (47) of entropy conservative fluxes. Indeed, for $f(u(w))$ as in (5),

$$\begin{aligned} & \int_0^1 f_h \circ u((1-s)w_- + sw_+) ds \\ &= \frac{1}{g} \int_0^1 \left(((1-s)w_{1-} + sw_{1+})((1-s)w_{2-} + sw_{2+}) + \frac{1}{2}((1-s)w_{2-} + sw_{2+})^3 \right) ds \\ &= \frac{1}{g} \left(\frac{w_{1+}w_{2+}}{3} + \frac{w_{1+}w_{2-}}{6} + \frac{w_{1-}w_{2+}}{6} + \frac{w_{1-}w_{2-}}{3} + \frac{w_{2+}^3}{8} + \frac{w_{2+}^2w_{2-}}{8} + \frac{w_{2+}w_{2-}^2}{8} + \frac{w_{2-}^3}{8} \right), \end{aligned} \quad (50)$$

and

$$\begin{aligned} & \int_0^1 f_{hv} \circ u((1-s)w_- + sw_+) ds \\ &= \frac{1}{g} \int_0^1 \left(\frac{1}{2}((1-s)w_{1-} + sw_{1+})^2 + \frac{3}{2}((1-s)w_{1-} + sw_{1+})((1-s)w_{2-} + sw_{2+})^2 \right. \\ & \quad \left. + \frac{5}{8}((1-s)w_{2-} + sw_{2+})^4 \right) ds \\ &= \frac{1}{24g} \left(4w_{1+}^2 + 4w_{1+}w_{1-} + 9w_{1+}w_{2+}^2 + 6w_{1+}w_{2+}w_{2-} + 3w_{1+}w_{2-}^2 + 4w_{1-}^2 + 3w_{1-}w_{2+}^2 \right. \\ & \quad \left. + 6w_{1-}w_{2+}w_{2-} + 9w_{1-}w_{2-}^2 + 3w_{2+}^4 + 3w_{2+}^3w_{2-} + 3w_{2+}^2w_{2-}^2 + 3w_{2+}w_{2-}^3 + 3w_{2-}^4 \right). \end{aligned} \quad (51)$$

Comparing this with the numerical fluxes $f_{a_1, a_2}^{\text{num}}$ (48), it can be seen that Tadmor's flux as above is recovered by setting $a_1 = a_2 = \frac{1}{3}$.

As proved by Fisher and Carpenter (2013, Theorem 3.2), a two-point entropy conservative flux as $f_{a_1, a_2}^{\text{num}}$ (47) can be used to construct a high-order spatial discretisation for diagonal-norm SBP operators with nodal basis including boundary nodes. Gassner, Winters, and Kopriva (2016b, Lemma 1) provided some examples for analogous split forms and numerical fluxes given by simple products of mean values. Analogously, using a diagonal-norm SBP derivative operator \underline{D} (i.e. an SBP derivative operator \underline{D} , where the corresponding norm / mass matrix \underline{M} is diagonal) with nodal basis including boundary nodes, the split form corresponding to the flux $f_{a_1, a_2}^{\text{num}}$ expressed via primitive variables (49) is for the h component given by

$$\begin{aligned} & [\text{VOL}_h^{a_1, a_2}]_i \\ &= \sum_{k=0}^p 2D_{i,k} \left(\frac{3-a_1}{8} (h_i v_i + h_k v_k) + \frac{1+a_1}{8} (h_i v_k + h_k v_i) \right. \\ & \quad \left. + \frac{a_1 + 3a_2 - 2}{16g} (v_i^3 - v_i^2 v_k - v_i v_k^2 + v_k^3) \right) \end{aligned} \quad (52)$$

$$\begin{aligned}
&= \sum_{k=0}^p D_{i,k} \left(\frac{3-a_1}{4} h_k v_k + \frac{1+a_1}{4} (h_i v_k + h_k v_i) + \frac{a_1+3a_2-2}{8g} (v_k^3 - v_i v_k^2 - v_i^2 v_k) \right) \\
&= \left[\frac{3-a_1}{4} \underline{\underline{D}} h v + \frac{1+a_1}{4} (\underline{\underline{h}} \underline{\underline{D}} v + v \underline{\underline{D}} h) + \frac{a_1+3a_2-2}{8g} (\underline{\underline{D}} v^3 - v \underline{\underline{D}} v^2 - v^2 \underline{\underline{D}} v) \right]_i.
\end{aligned}$$

Here, some summands have been dropped, because the derivative is exact for constants, i.e. $\underline{\underline{D}} \underline{\underline{1}} = 0$, resulting in $\sum_{k=0}^p D_{i,k} = 0$. The first two terms form a consistent discretisation of $\partial_x(hv) = (\partial_x h)v + h(\partial_x v)$ for smooth solutions. The third term is consistently zero, since $\partial_x v^3 = (\partial_x v^2)v + v^2(\partial_x v)$ by the product for smooth solutions.

Similarly, the hv component can be computed via

$$\begin{aligned}
&[\text{VOL}_{hv}^{a_1, a_2}]_i \tag{53} \\
&= \sum_{k=0}^p 2D_{i,k} \left(\frac{1+a_1}{8} g (h_i^2 + h_k^2) + \frac{1-a_1}{4} g h_i h_k - \frac{2a_1+3a_2-5}{16} (h_i v_i^2 + h_k v_k^2) \right. \\
&\quad + \frac{2a_1+3a_2-1}{16} (h_i v_k^2 + h_k v_i^2) + \frac{h_i v_i v_k}{4} + \frac{h_k v_i v_k}{4} \\
&\quad \left. + \frac{a_1+3a_2-2}{32g} (v_i^4 - 2v_i^2 v_k^2 + v_k^4) \right) \\
&= \sum_{k=0}^p D_{i,k} \left(\frac{1+a_1}{4} g h_k^2 + \frac{1-a_1}{2} g h_i h_k - \frac{2a_1+3a_2-5}{8} h_k v_k^2 + \frac{2a_1+3a_2-1}{8} (h_i v_k^2 + h_k v_i^2) \right. \\
&\quad \left. + \frac{h_i v_i v_k}{2} + \frac{h_k v_i v_k}{2} + \frac{a_1+3a_2-2}{16g} (v_k^4 - 2v_i^2 v_k^2) \right) \\
&= \left[\frac{1+a_1}{4} g \underline{\underline{D}} h^2 + \frac{1-a_1}{2} g \underline{\underline{h}} \underline{\underline{D}} h - \frac{2a_1+3a_2-5}{8} \underline{\underline{D}} h v^2 + \frac{2a_1+3a_2-1}{8} (\underline{\underline{h}} \underline{\underline{D}} v^2 + v^2 \underline{\underline{D}} h) \right. \\
&\quad \left. + \frac{1}{2} (\underline{\underline{h}} v \underline{\underline{D}} v + v \underline{\underline{D}} h v) + \frac{a_1+3a_2-2}{16g} (\underline{\underline{D}} v^4 - 2v^2 \underline{\underline{D}} v^2) \right]_i,
\end{aligned}$$

where again $\sum_{k=0}^p D_{i,k} = 0$ has been used. The first two terms form a consistent discretisation of $\frac{1}{2}g\partial_x h^2$, the three following terms are consistent with $\partial_x hv^2$ and the last two terms are consistently zero. This is summed up in

Lemma 3. *The volume terms (52), (53) are consistent volume discretisations of the shallow water equations with vanishing bottom topography (1) and yield an entropy conservative volume discretisation using diagonal-norm SBP operators with nodal bases including boundary nodes.*

The entropy conservation follows from the general result of Fisher and Carpenter (2013, Theorem 3.2) and will also be investigated in more detail for Gauß nodes and other SBP bases in section 6.

5 Adding well-balanced source discretisations

In this section, the discretisation of the source term $gh\partial_x b$ in the shallow water equations (1) will be investigated. It should be *consistent*, *stable*, and *well-balanced*, if combined with the remaining semidiscretisation derived in the previous section.

5.1 Connections between finite volume and SBP SAT schemes

A general semidiscretisation of a conservation law

$$\partial_t u + \partial_x f(u) = 0 \tag{54}$$

with a polynomial SBP method using the notation of Ranocha, Öffner, and Sonar (2016) can be written as

$$\partial_t \underline{u} = -\underline{\text{VOL}} + \underline{\text{SURF}} - \underline{M}^{-1} \underline{R}^T \underline{B} \underline{f}^{\text{num}}, \quad (55)$$

where $\underline{\text{VOL}}$ contains the volume terms consistent with $\partial_x f(u)$, possibly using some split form, f^{num} is the numerical (surface) flux, and $\underline{\text{SURF}}$ contains additional surface terms, consistent with the difference of the flux values $f(u)$ at the boundaries, i.e. $\underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} f$ in the simplest case, but additional terms may also appear, especially if a nodal basis without boundary nodes is used, see also section 6.

If the polynomial degree p is set to zero, the volume terms vanish, since the derivative is exact for constants, i.e. $\underline{D} \underline{1} = 0$. Additionally, since the extra surface terms $\underline{\text{SURF}}$ are a consistent evaluation of the difference of boundary values, they vanish, too, because this difference is zero for constants. Therefore, this method reduces to a simple finite volume method. If the cell i is of size Δx and the flux f^{num} between the cells i and k is denoted as $f_{i,k}^{\text{num}}$, the FV method can be written as

$$\partial_t u_i = -\frac{1}{\Delta x} \left(f_{i,i+1}^{\text{num}} - f_{i,i-1}^{\text{num}} \right), \quad (56)$$

and is determined solely by the numerical flux f^{num} used at the boundaries.

On the other hand, using the theory of Fisher and Carpenter (2013), a finite volume method with entropy conservative flux f^{num} can be used to construct the volume terms $\underline{\text{VOL}}$, if a nodal SBP basis including boundary nodes is used. In this case, since the evaluation at the boundary is exact and commutes with nonlinear operations, the surface terms are simply $\underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} f$.

This strong correspondence between SBP schemes and FV methods will be used in the following sections to extend results from one area to the other and vice versa.

5.2 Results of the FV setting

Setting the polynomial degree p of the SBP SAT semidiscretisation (21) of Wintermeyer, Winters, Gassner, and Kopriva (2016) to zero results in the following FV scheme for the time derivative of hv

$$\begin{aligned} \partial_t (hv)_i = & -\frac{1}{\Delta x} \left(f_{hv}^{\text{num}}(u_i, u_{i+1}) - f_{hv}^{\text{num}}(u_{i-1}, u_i) \right. \\ & \left. + \frac{1}{2} g \frac{h_{i-1} + h_i}{2} (b_i - b_{i-1}) + \frac{1}{2} g \frac{h_i + h_{i+1}}{2} (b_{i+1} - b_i) \right). \end{aligned} \quad (57)$$

This is the same FV scheme as the entropy conservative and well-balanced one proposed by Fjordholm, Mishra, and Tadmor (2011). Using the notation

$$\{a\}_{i,k} = \frac{a_i + a_k}{2}, \quad [a]_{i,k} = a_k - a_i, \quad f_{hv,i,k}^{\text{num}} = f_{hv}^{\text{num}}(u_i, u_k), \quad (58)$$

this can be rewritten as

$$\begin{aligned} \partial_t (hv)_i = & -\frac{1}{\Delta x} \left(f_{hv,i,i+1}^{\text{num}} - f_{hv,i-1,i}^{\text{num}} \right) - \frac{1}{2\Delta x} g \left(\{h\}_{i,i+1} [b]_{i,i+1} + \{h\}_{i-1,i} [b]_{i-1,i} \right) \\ = & -\frac{1}{\Delta x} \left(\left[f_{hv,i,i+1}^{\text{num}} + \frac{g}{2} \{h\}_{i,i+1} [b]_{i,i+1} \right] - \left[f_{hv,i-1,i}^{\text{num}} + \frac{g}{2} \{h\}_{i-1,i} [b]_{i-1,i} \right] \right), \end{aligned} \quad (59)$$

since $[b]_{i-1,i} = -[b]_{i,i-1}$, and both the mean $\{h\}_{i,j}$ and the numerical flux $f_{i,j}^{\text{num}}$ are symmetric with respect to the indices i, j . Therefore, this scheme can be interpreted as a finite volume method with extended flux

$$f_{hv,i,i+1}^{\text{num}} + \frac{1}{2} g \{h\}_{i,i+1} [b]_{i,i+1}, \quad f_{hv,i,k}^{\text{num}} = \{h\}_{i,k} \{v\}_{i,k}^2 + \frac{1}{2} g \{h^2\}_{i,k}, \quad (60)$$

instead of $f_{hv,i,i+1}^{\text{num}}$, in order to include the source term. This corresponds to the terms appearing in the semidiscretisation (21) of Wintermeyer, Winters, Gassner, and Kopriva (2016), consisting of a numerical flux and the jump terms at the boundary.

Thus, in a FV setting, the source term can be incorporated into the numerical flux, resulting in an extended numerical flux

$$f_{i,k}^{\text{num,ext}} = f_{i,k}^{\text{num}} + S_{i,k}, \quad (61)$$

where $f_{i,k}^{\text{num}}$ is a usual symmetric numerical flux of the problem without source terms and $S_{i,k}$ described the source terms and is not necessarily symmetric.

Rewriting the FV evolution equation (56) by adding $f_i - f_i = 0$ (motivated by the form of SBP SAT methods) and using extended numerical fluxes yields

$$\partial_t u_i = -\frac{1}{\Delta x} \left(\left(f_{i,i+1}^{\text{num,ext}} - f_i \right) - \left(f_{i,i-1}^{\text{num,ext}} - f_i \right) \right). \quad (62)$$

Therefore, the rate of change of the entropy U can be calculated as

$$\begin{aligned} \partial_t U_i &= w_i \cdot \partial_t u_i = -\frac{1}{\Delta x} \left(w_i \cdot \left(f_{i,i+1}^{\text{num,ext}} - f_i \right) - w_i \cdot \left(f_{i,i-1}^{\text{num,ext}} - f_i \right) \right) \\ &= -\frac{1}{\Delta x} \left(\left[w_i \cdot \left(f_{i,i+1}^{\text{num,ext}} - f_i \right) + F_i \right] - \left[w_i \cdot \left(f_{i,i-1}^{\text{num,ext}} - f_i \right) + F_i \right] \right). \end{aligned} \quad (63)$$

Thus, adding the contributions of the right hand side of cell i and the left hand side of cell $i+1$ yields after multiplication with Δx

$$\begin{aligned} &\left[w_{i+1} \cdot \left(f_{i+1,i}^{\text{num,ext}} - f_{i+1} \right) + F_{i+1} \right] - \left[w_i \cdot \left(f_{i,i+1}^{\text{num,ext}} - f_i \right) + F_i \right] \\ &= (w_{i+1} - w_i) \cdot f_{i,i+1}^{\text{num}} - \left(\underbrace{[w_{i+1} \cdot f_{i+1} + F_{i+1}]}_{=\psi_{i+1}} - \underbrace{[w_i \cdot f_i + F_i]}_{=\psi_i} \right) + w_{i+1} \cdot S_{i+1,i} - w_i \cdot S_{i,i+1}, \end{aligned} \quad (64)$$

where the extended flux $f^{\text{num,ext}}$ (61) has been inserted and the symmetry of the numerical flux f^{num} has been used.

Assume now that the numerical flux $f_{i,i+1}^{\text{num}}$ is chosen as an entropy conservative one, fulfilling $\llbracket w \rrbracket_{i,i+1} \cdot f_{i,i+1}^{\text{num}} = \llbracket \psi \rrbracket_{i,i+1}$ (28), for vanishing bottom topography $b \equiv 0$. Here, the entropy variables are $w = \left(gh - \frac{1}{2}v^2, v \right)^T$, since $b \equiv 0$. In the general case, the entropy variables are $w = \left(g(h+b) - \frac{1}{2}v^2, v \right)^T$, resulting in

$$\llbracket w \rrbracket_{i,i+1} \cdot f_{i,i+1}^{\text{num}} = \llbracket \psi \rrbracket_{i,i+1} + g f_{h,i+1}^{\text{num}} \llbracket b \rrbracket_{i,i+1}. \quad (65)$$

Thus, the contribution of one boundary to the rate of change of the entropy (64) is

$$g f_{h,i+1}^{\text{num}} \llbracket b \rrbracket_{i,i+1} + w_{i+1} \cdot S_{i+1,i} - w_i \cdot S_{i,i+1} \stackrel{!}{=} 0. \quad (66)$$

This proves

Lemma 4 (cf. Lemma 2.1 of Fjordholm, Mishra, and Tadmor (2011) and Appendix B.1 of Wintermeyer, Winters, Gassner, and Kopriva (2016)). *If the source discretisation $S_{i,k}$ in the extended numerical flux (61) is chosen such that the expression (66) is zero for an entropy conservative numerical flux f^{num} (28) for the shallow water equations with vanishing bottom topography $b \equiv 0$, then the resulting scheme is entropy conservative for general bottom topography.*

The source discretisation of Fjordholm, Mishra, and Tadmor (2011) results in the extended numerical flux (60) with source terms $S_{i,k} = \frac{1}{2}g \llbracket h \rrbracket_{i,k} \llbracket b \rrbracket_{i,k}$. Thus, inserting this and their numerical flux $f_{h,i,k}^{\text{num}} = \llbracket h \rrbracket_{i,k} \llbracket v \rrbracket_{i,k}$ into (66) results in

$$\begin{aligned} &g \llbracket h \rrbracket_{i,i+1} \llbracket v \rrbracket_{i,i+1} \llbracket b \rrbracket_{i,i+1} + v_{i+1} \frac{g}{2} \llbracket h \rrbracket_{i,i+1} \llbracket b \rrbracket_{i+1,i} - v_i \frac{g}{2} \llbracket h \rrbracket_{i+1,i} \llbracket b \rrbracket_{i,i+1} \\ &= g \llbracket h \rrbracket_{i,i+1} \llbracket v \rrbracket_{i,i+1} \llbracket b \rrbracket_{i,i+1} - g \llbracket h \rrbracket_{i,i+1} \frac{v_i + v_{i+1}}{2} \llbracket b \rrbracket_{i,i+1} = 0, \end{aligned} \quad (67)$$

fulfilling the condition of Lemma 4.

The extended flux (61) results in a well-balanced FV scheme (56), if the right hand side is zero for vanishing discharge $h\nu$ and constant total height $h + b$. This is fulfilled by the numerical flux (23) of Fjordholm, Mishra, and Tadmor (2011); Gassner, Winters, and Kopriva (2016a). To see this, the calculation

$$\begin{aligned} \{h^2\}_{i,i+1} - \{h^2\}_{i,i-1} &= \frac{h_i^2 + h_{i+1}^2}{2} - \frac{h_i^2 + h_{i-1}^2}{2} \\ &= \frac{h_i + h_{i+1}}{2}(h_{i+1} - h_i) - \frac{h_i + h_{i-1}}{2}(h_{i-1} - h_i) \\ &= \{h\}_{i,i+1} [h]_{i,i+1} - \{h\}_{i,i-1} [h]_{i,i-1} \end{aligned} \quad (68)$$

can be used, resulting in

$$\begin{aligned} f_{hv,i,i+1}^{\text{num,ext}} - f_{hv,i,i-1}^{\text{num,ext}} &= \left(\frac{g}{2} \{h^2\}_{i,i+1} + \frac{g}{2} \{h\}_{i,i+1} [b]_{i,i+1} \right) - \left(\frac{g}{2} \{h^2\}_{i,i-1} + \frac{g}{2} \{h\}_{i,i-1} [b]_{i,i-1} \right) \\ &= \frac{g}{2} \left(\{h\}_{i,i+1} [h]_{i,i+1} + \{h\}_{i,i+1} [b]_{i,i+1} - \{h\}_{i,i-1} [h]_{i,i-1} - \{h\}_{i,i-1} [b]_{i,i-1} \right) = 0, \end{aligned} \quad (69)$$

since the velocity vanishes and $[h + b] = 0$ for the lake-at-rest initial condition. Thus, this FV scheme is well-balanced and entropy conservative.

5.3 Results of the SBP SAT setting

If the bottom topography b is continuous across boundaries of the cells, the numerical surface flux corresponds to a finite volume flux with constant bottom topography, and the only source contributions can be found in the volume terms. In this setting of Gassner, Winters, and Kopriva (2016a), the relevant volume terms for the lake-at-rest condition in the semidiscretisation (21) are ($\underline{h\nu} = 0 = \underline{\nu}$)

$$\text{VOL}_{h\nu} = g\underline{h} \underline{D} \underline{h} + g\underline{h} \underline{D} \underline{b} = g\underline{h} \underline{D} (\underline{h} + \underline{b}) = 0, \quad (70)$$

since the derivative is exact for constants. Thus, this part of the scheme is well-balanced.

The entropy conservation can be seen using the SBP property as described in the following section 6 or by Gassner, Winters, and Kopriva (2016a), or by an investigation of the numerical volume flux (24) yielding the split form semidiscretisation (21) as done by Wintermeyer, Winters, Gassner, and Kopriva (2016), using the results of Fisher and Carpenter (2013).

5.4 Combining the results into source discretisations of the entropy conservative fluxes

Transferring the results of the FV setting to the volume terms in an SBP semidiscretisation, inserting the extended flux (60) in the flux differencing form of Fisher and Carpenter (2013, Theorem 3.2), the volume terms $\frac{1}{2}g\partial_x h^2 + gh\partial_x b$ can be discretised as

$$\begin{aligned} &\sum_{k=0}^p 2D_{i,k} \left(\frac{1}{2}g \{h^2\}_{i,k} + \frac{1}{2}g \{h\}_{i,k} [b]_{i,k} \right) \\ &= \frac{1}{2}g \sum_{k=0}^p D_{i,k} \left(h_i^2 + h_k^2 + (h_i + h_k)(b_k - b_i) \right) \\ &= \frac{1}{2}gh_i^2 \underbrace{\sum_{k=0}^p D_{i,k}}_{=0} + \frac{1}{2}g \sum_{k=0}^p D_{i,k} (h_k(h_k + b_k) + h_i b_k - h_k b_i) - \frac{1}{2}gh_i b_i \underbrace{\sum_{k=0}^p D_{i,k}}_{=0} \\ &= \frac{1}{2}g \left[\underline{D} (\underline{h} + \underline{b}) \underline{h} + \underline{h} \underline{D} \underline{b} - \underline{b} \underline{D} \underline{h} \right]_i. \end{aligned} \quad (71)$$

Therefore, if $h + b \equiv \text{const}$, $\left(\underline{\underline{h}} + \underline{\underline{b}}\right) = (h + b)\underline{\underline{1}}$, and

$$\begin{aligned} \underline{\underline{D}} \left(\underline{\underline{h}} + \underline{\underline{b}} \right) \underline{\underline{h}} + \underline{\underline{h}} \underline{\underline{D}} \underline{\underline{b}} - \underline{\underline{b}} \underline{\underline{D}} \underline{\underline{h}} &= \left(\underline{\underline{h}} + \underline{\underline{b}} \right) \underline{\underline{D}} \underline{\underline{h}} + \underline{\underline{h}} \underline{\underline{D}} \underline{\underline{b}} - \underline{\underline{b}} \underline{\underline{D}} \underline{\underline{h}} \\ &= \underline{\underline{h}} \underline{\underline{D}} (\underline{\underline{h}} + \underline{\underline{b}}) = 0. \end{aligned} \quad (72)$$

Thus, this discretisation is also well-balanced. It is entropy conservative, since the numerical flux is entropy conservative (Fisher and Carpenter, 2013, Theorem 3.2) (at least for vanishing b). This will also be proven in the more general setting of section 6.

On the other hand, the discretisation of the volume terms $\frac{1}{2}g\partial_x h^2 + gh\partial_x b$ by Gassner, Winters, and Kopriva (2016a) can be expressed using

$$\left[gh \underline{\underline{D}} \underline{\underline{b}} \right]_i = g \sum_{k=0}^p h_i D_{i,k} b_k = g \sum_{k=0}^p h_i D_{i,k} (b_k - b_i) = \sum_{k=0}^p 2D_{i,k} \frac{1}{2} g h_i \llbracket b \rrbracket_{i,k}, \quad (73)$$

since $\underline{\underline{D}} \underline{\underline{1}} = 0$, as

$$\begin{aligned} &\sum_{k=0}^p 2D_{i,k} \left(g \llbracket h \rrbracket_{i,k}^2 - \frac{1}{2} g \llbracket h^2 \rrbracket_{i,k} + \frac{1}{2} g h_i \llbracket b \rrbracket_{i,k} \right) \\ &= g \sum_{k=0}^p 2D_{i,k} \left(\left(\frac{h_i + h_k}{2} \right)^2 - \frac{1}{2} \frac{h_i^2 + h_k^2}{2} + \frac{1}{2} h_i (b_k - b_i) \right) \\ &= g \sum_{k=0}^p D_{i,k} (h_i h_k + h_i (b_k - b_i)) = g h_i \underbrace{\sum_{k=0}^p D_{i,k} (h_k + b_k)}_{=[\underline{\underline{D}}(\underline{\underline{h}}+\underline{\underline{b}})]_i} - g h_i b_i \underbrace{\sum_{k=0}^p D_{i,k}}_{=[\underline{\underline{D}} \underline{\underline{1}}]_i=0}. \end{aligned} \quad (74)$$

Again, this computation translates the flux difference form using f^{vol} to the split operator form in (21) as shown by Gassner, Winters, and Kopriva (2016b, Lemma 1). Analogously to (60), this yields an extended flux

$$f_{hv_{i,k}}^{\text{vol}} + \frac{1}{2} g h_i \llbracket b \rrbracket_{i,k}, \quad f_{hv_{i,k}}^{\text{vol}} = \llbracket hv \rrbracket_{i,k} \llbracket v \rrbracket_{i,k} + g \llbracket h \rrbracket_{i,k}^2 - \frac{1}{2} g \llbracket h^2 \rrbracket_{i,k}. \quad (75)$$

In the same way the entropy conservative numerical fluxes (23) and (24) can be combined to get the one-parameter family of entropy conservative fluxes f_α^{num} (36) for the shallow water equations with vanishing bottom topography $b \equiv 0$, these extended numerical fluxes can be combined to get entropy conservative extended numerical fluxes for the shallow water equations with general bottom topography. This proves

Lemma 5. *The one-parameter family*

$$\begin{aligned} f_{h_{i,k}}^\alpha &= \frac{1-\alpha}{2} \llbracket hv \rrbracket_{i,k} + \frac{1+\alpha}{2} \llbracket h \rrbracket_{i,k} \llbracket v \rrbracket_{i,k} \\ &\stackrel{(37)}{=} \frac{3-\alpha}{8} (h_i v_i + h_k v_k) + \frac{1+\alpha}{8} (h_i v_k + h_k v_i), \\ f_{hv_{i,k}}^\alpha &= \frac{1-\alpha}{2} \llbracket hv \rrbracket_{i,k} \llbracket v \rrbracket_{i,k} + \frac{1+\alpha}{2} \llbracket h \rrbracket_{i,k} \llbracket v \rrbracket_{i,k}^2 + g \frac{\alpha}{2} \llbracket h^2 \rrbracket_{i,k} \\ &\quad + \frac{1-\alpha}{2} g \llbracket h \rrbracket_{i,k}^2 + \frac{1-\alpha}{4} g h_i \llbracket b \rrbracket_{i,k} + \frac{1+\alpha}{4} g \llbracket h \rrbracket_{i,k} \llbracket b \rrbracket_{i,k} \\ &\stackrel{(37)}{=} \frac{1+\alpha}{8} g (h_i^2 + h_k^2) + \frac{1-\alpha}{4} g h_i h_k + \frac{3-\alpha}{16} (h_i v_i^2 + h_k v_k^2) + \frac{1+\alpha}{16} (h_i v_k^2 + h_k v_i^2) \\ &\quad + \frac{h_i v_i v_k}{4} + \frac{h_k v_i v_k}{4} + \frac{1}{4} g \left(\frac{3-\alpha}{2} h_i + \frac{1+\alpha}{2} h_k \right) (b_k - b_i), \end{aligned} \quad (76)$$

is a family of entropy conservative and well-balanced extended numerical fluxes (including contributions of the source term $gh\partial_x b$) for the shallow water equations (1) with general bottom topography b .

Again, the extended fluxes (75) and (60) can be obtained by setting $\alpha = -1$ and $\alpha = 1$, respectively.

Comparing the two-parameter family of numerical fluxes $f_{a_1, a_2}^{\text{num}}$ (49) with the one-parameter family f_α^{num} (37), the additional parameter a_2 contributes only to terms containing the velocity v . Thus, it is irrelevant for well-balancing, since these terms vanish for the lake-at-rest initial condition.

However, the source discretisation $S_{i,k}$ in the extended numerical flux (61) has to be adapted to the additional terms with a_2 in order to fulfil the condition (66) of Lemma 4. Since the two-parameter flux (49) for h contains an additional term $\frac{a_1+3a_2-2}{16g} (v_+^3 - v_+^2 v_- - v_+ v_-^2 + v_-^3)$ compared to the one-parameter flux (37), the new source term $S_{i,k}$ can be written as the sum of the source term $\frac{1}{4}g \left(\frac{3-\alpha}{2} h_i + \frac{1+\alpha}{2} h_k \right) (b_k - b_i)$ for the one-parameter flux (37) and an additional source term $\tilde{S}_{i,k}$, obeying

$$\begin{aligned} & g f_{h_{i,k}}^{\text{num}} [[b]]_{i,k} + w_k \cdot S_{k,i} - w_i \cdot S_{i,k} \\ &= \frac{a_1 + 3a_2 - 2}{16} \left(v_i^3 - v_i^2 v_k - v_i v_k^2 + v_k^3 \right) (b_k - b_i) + v_k \tilde{S}_{k,i} - v_i \tilde{S}_{i,k} \stackrel{!}{=} 0. \end{aligned} \quad (77)$$

This can be rewritten using

$$v_i^3 - v_i^2 v_k - v_i v_k^2 + v_k^3 = (v_i - v_k)(v_i^2 - v_k^2) = (v_i + v_k)(v_i - v_k)^2. \quad (78)$$

Thus, choosing $\tilde{S}_{i,k} = \frac{a_1+3a_2-2}{16} (v_k - v_i)^2 (b_k - b_i)$ results in the desired equality

$$\begin{aligned} & \frac{a_1 + 3a_2 - 2}{16} \left(v_i^3 - v_i^2 v_k - v_i v_k^2 + v_k^3 \right) (b_k - b_i) + v_k \tilde{S}_{k,i} - v_i \tilde{S}_{i,k} \\ &= \frac{a_1 + 3a_2 - 2}{16} \left((v_i + v_k)(v_i - v_k)^2 (b_k - b_i) + v_k (v_i - v_k)^2 (b_i - b_k) - v_i (v_k - v_i)^2 (b_k - b_i) \right) \\ &= 0. \end{aligned} \quad (79)$$

This proves

Lemma 6. *The two-parameter family*

$$\begin{aligned} f_{h_{i,k}}^{a_1, a_2} &= \frac{3 - a_1}{8} (h_i v_i + h_k v_k) + \frac{1 + a_1}{8} (h_i v_k + h_k v_i) + \frac{a_1 + 3a_2 - 2}{16g} \left(v_i^3 - v_i^2 v_k - v_i v_k^2 + v_k^3 \right), \\ f_{hv_{i,k}}^{a_1, a_2} &= \frac{1 + a_1}{8} g \left(h_i^2 + h_k^2 \right) + \frac{1 - a_1}{4} g h_i h_k - \frac{2a_1 + 3a_2 - 5}{16} \left(h_i v_i^2 + h_k v_k^2 \right) \\ &\quad + \frac{2a_1 + 3a_2 - 1}{16} \left(h_i v_k^2 + h_k v_i^2 \right) + \frac{h_i v_i v_k}{4} + \frac{h_k v_i v_k}{4} + \frac{a_1 + 3a_2 - 2}{32g} \left(v_i^4 - 2v_i^2 v_k^2 + v_k^4 \right) \\ &\quad + \frac{1}{4} g \left(\frac{3 - a_1}{2} h_i + \frac{1 + a_1}{2} h_k \right) (b_k - b_i) + \frac{a_1 + 3a_2 - 2}{16} (v_k - v_i)^2 (b_k - b_i), \end{aligned} \quad (80)$$

is a family of entropy conservative and well-balanced extended numerical fluxes (including contributions of the source term $gh\partial_x b$) for the shallow water equations (1) with general bottom topography b .

Again, the one-parameter family (76) of Lemma 5 is given by the special choice $a_2 = \frac{2-a_1}{3}$ and $a_1 = \alpha$.

The volume terms corresponding to the two-parameter family of fluxes (80) are given by

$$\begin{aligned} \text{VOL}_h^{a_1, a_2} &\stackrel{(52)}{=} \frac{3 - a_1}{4} \underline{\underline{D}} \underline{\underline{h}} v + \frac{1 + a_1}{4} \left(\underline{\underline{h}} \underline{\underline{D}} v + \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{h}} \right) + \frac{a_1 + 3a_2 - 2}{8g} \left(\underline{\underline{D}} v^3 - \underline{\underline{v}} \underline{\underline{D}} v^2 - v^2 \underline{\underline{D}} v \right), \\ \text{VOL}_{hv}^{a_1, a_2} &\stackrel{(53)}{=} \frac{1 + a_1}{4} g \underline{\underline{D}} \underline{\underline{h}}^2 + \frac{1 - a_1}{2} g \underline{\underline{h}} \underline{\underline{D}} \underline{\underline{h}} - \frac{2a_1 + 3a_2 - 5}{8} \underline{\underline{D}} \underline{\underline{h}} v^2 \\ &\quad + \frac{2a_1 + 3a_2 - 1}{8} \left(\underline{\underline{h}} \underline{\underline{D}} v^2 + \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{h}} \right) + \frac{1}{2} \left(\underline{\underline{h}} v \underline{\underline{D}} v + \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{h}} v \right) \\ &\quad + \frac{a_1 + 3a_2 - 2}{16g} \left(\underline{\underline{D}} v^4 - 2\underline{\underline{v}} \underline{\underline{D}} v^2 \right) + \frac{3 - a_1}{4} g \underline{\underline{h}} \underline{\underline{D}} \underline{\underline{b}} + \frac{1 + a_1}{4} g \left(\underline{\underline{D}} \underline{\underline{h}} \underline{\underline{b}} - \underline{\underline{b}} \underline{\underline{D}} \underline{\underline{h}} \right) \\ &\quad + \frac{a_1 + 3a_2 - 2}{8} \left(\underline{\underline{D}} \underline{\underline{b}} v^2 - \underline{\underline{b}} \underline{\underline{D}} v^2 - 2\underline{\underline{v}} \underline{\underline{D}} \underline{\underline{b}} v + \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{b}} + 2\underline{\underline{b}} v \underline{\underline{D}} v \right), \end{aligned} \quad (81)$$

where

$$\begin{aligned}
& \sum_{k=0}^p 2D_{i,k} \left(\frac{1}{4}g \left(\frac{3-a_1}{2}h_i + \frac{1+a_1}{2}h_k \right) (b_k - b_i) + \frac{a_1+3a_2-2}{16}(v_k - v_i)^2(b_k - b_i) \right) \quad (82) \\
&= \sum_{k=0}^p D_{i,k} \left(\frac{3-a_1}{4}gh_i b_k + \frac{1+a_1}{4}gh_k(b_k - b_i) \right. \\
&\quad \left. + \frac{a_1+3a_2-2}{8} \left(v_k^2 b_k - 2v_i v_k b_k + v_i^2 b_k - v_k^2 b_i + 2v_i b_i v_k \right) \right) \\
&= \left[\frac{3-a_1}{4}g \underline{\underline{h}} \underline{\underline{D}} \underline{\underline{b}} + \frac{1+a_1}{4}g \left(\underline{\underline{D}} \underline{\underline{h}} \underline{\underline{b}} - \underline{\underline{b}} \underline{\underline{D}} \underline{\underline{h}} \right) \right. \\
&\quad \left. + \frac{a_1+3a_2-2}{8} \left(\underline{\underline{D}} \underline{\underline{b}} \underline{\underline{v}}^2 - 2\underline{\underline{v}} \underline{\underline{D}} \underline{\underline{b}} \underline{\underline{v}} + \underline{\underline{v}}^2 \underline{\underline{D}} \underline{\underline{b}} - \underline{\underline{b}} \underline{\underline{D}} \underline{\underline{v}}^2 + 2\underline{\underline{b}} \underline{\underline{v}} \underline{\underline{D}} \underline{\underline{v}} \right) \right]_i
\end{aligned}$$

has been used. The corresponding surface terms using nodal bases including boundary nodes are simply given by

$$\underline{\underline{\text{SURF}}}_h = \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} \underline{\underline{h}} \underline{\underline{v}}, \quad \underline{\underline{\text{SURF}}}_{hv} = \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \left(\underline{\underline{R}} \underline{\underline{h}} \underline{\underline{v}}^2 + \frac{1}{2}g \underline{\underline{R}} \underline{\underline{h}}^2 \right). \quad (83)$$

The numerical fluxes used for the volume discretisation and as surface flux may be combined arbitrarily, as done by Gassner, Winters, and Kopriva (2016b); Wintermeyer, Winters, Gassner, and Kopriva (2016), where they have used $f^{-1,1}$ as volume flux and $f^{1,\frac{1}{3}}$ as surface flux. Thus, a general semidiscretisation is of the form

$$\begin{aligned}
\partial_t \underline{\underline{h}} &= -\underline{\underline{\text{VOL}}}_h^{a_1, a_2} + \underline{\underline{\text{SURF}}}_h - \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{f}}_h^{b_1, b_2}, \\
\partial_t \underline{\underline{h}} \underline{\underline{v}} &= -\underline{\underline{\text{VOL}}}_{hv}^{a_1, a_2} + \underline{\underline{\text{SURF}}}_{hv} - \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{f}}_{hv}^{b_1, b_2},
\end{aligned} \quad (84)$$

with $a_1, a_2, b_1, b_2 \in \mathbb{R}$. This yields

Lemma 7. *For parameters $a_1, a_2, b_1, b_2 \in \mathbb{R}$, the semidiscretisation (84) with volume terms (81) and surface terms (83) using SBP operators on a nodal basis including boundary nodes*

1. *conserves the mass in general and the discharge if the bottom topography is constant,*
2. *conserves the entropy / total energy,*
3. *handles the lake-at-rest condition correctly.*

That is, the semidiscretisations (84) are conservative, stable (entropy conservative), and well-balanced.

6 Extension to general SBP bases

In this section, an extension of the previous result to a nodal DG method using Gauß nodes instead of Lobatto nodes or more general SBP bases will be investigated.

Although the volume terms (81) have been derived in section 5 with the assumption of a diagonal-norm SBP basis including boundary nodes, they can be easily transferred to the setting of a general SBP basis. If the multiplication operators are self-adjoint with respect to the scalar product induced by $\underline{\underline{M}}$, e.g. for a nodal basis with diagonal mass matrix, then the same volume terms (81) can be used. Otherwise, some multiplication operators $\underline{\underline{a}}$ have to be replaced by their $\underline{\underline{M}}$ -adjoints $\underline{\underline{a}}^* = \underline{\underline{M}}^{-1} \underline{\underline{a}}^T \underline{\underline{M}}$, as proposed by Ranocha, Öffner, and Sonar (2015). This results

in the volume terms

$$\begin{aligned}
\text{VOL}_h^{a_1, a_2} &= \frac{3-a_1}{4} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} + \frac{1+a_1}{4} \left(\underline{\underline{h}}^* \underline{\underline{D}} \underline{\underline{v}} + \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{h}} \right) + \frac{a_1+3a_2-2}{8g} \left(\underline{\underline{D}} \underline{\underline{v}}^3 - \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{v}}^2 - \underline{\underline{v}}^{2*} \underline{\underline{D}} \underline{\underline{v}} \right), \\
\text{VOL}_{hv}^{a_1, a_2} &= \frac{1+a_1}{4} g \underline{\underline{D}} \underline{\underline{h}}^2 + \frac{1-a_1}{2} g \underline{\underline{h}}^* \underline{\underline{D}} \underline{\underline{h}} - \frac{2a_1+3a_2-5}{8} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}}^2 \\
&\quad + \frac{2a_1+3a_2-1}{8} \left(\underline{\underline{h}}^* \underline{\underline{D}} \underline{\underline{v}}^2 + \underline{\underline{v}}^{2*} \underline{\underline{D}} \underline{\underline{h}} \right) + \frac{1}{2} \left(\underline{\underline{h}} \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{v}} + \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} \right) \\
&\quad + \frac{a_1+3a_2-2}{16g} \left(\underline{\underline{D}} \underline{\underline{v}}^4 - 2 \underline{\underline{v}}^{2*} \underline{\underline{D}} \underline{\underline{v}}^2 \right) + \frac{3-a_1}{4} g \underline{\underline{h}}^* \underline{\underline{D}} \underline{\underline{b}} + \frac{1+a_1}{4} g \left(\underline{\underline{D}} \underline{\underline{h}} \underline{\underline{b}} - \underline{\underline{b}}^* \underline{\underline{D}} \underline{\underline{h}} \right) \\
&\quad + \frac{a_1+3a_2-2}{8} \left(\underline{\underline{D}} \underline{\underline{b}} \underline{\underline{v}}^2 - \underline{\underline{b}}^* \underline{\underline{D}} \underline{\underline{v}}^2 - 2 \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{b}} \underline{\underline{v}} + \underline{\underline{v}}^{2*} \underline{\underline{D}} \underline{\underline{b}} + 2 \underline{\underline{b}} \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{v}} \right).
\end{aligned} \tag{85}$$

However, the surface terms (83) also have to be adapted to a general basis. Often, the split form of the volume terms is described as some correction for the product rule that does not hold discretely. However, as described by Ranocha, Öffner, and Sonar (2016), it is the multiplication that is not correct on a discrete level, resulting in an invalid product rule. Moreover, if no boundary nodes are included in the basis, this inexactness also has to be compensated in the surface terms. Thus, in the same spirit as the split form of the volume terms can be seen as corrections to inexact multiplication, some kind of correction has to be used for the interpolation to the boundaries.

Investigating *conservation* (across elements), the time derivatives of the conserved variables (84) are multiplied with $\underline{\underline{1}}^T \underline{\underline{M}}$, corresponding to integration over an element. This yields for the volume terms (85)

$$\begin{aligned}
\underline{\underline{1}}^T \underline{\underline{M}} \text{VOL}_h^{a_1, a_2} &= \frac{3-a_1}{4} \underline{\underline{1}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} + \frac{1+a_1}{4} \underline{\underline{1}}^T \underline{\underline{M}} \left(\underline{\underline{h}}^* \underline{\underline{D}} \underline{\underline{v}} + \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{h}} \right) \\
&\quad + \frac{a_1+3a_2-2}{8g} \underline{\underline{1}}^T \underline{\underline{M}} \left(\underline{\underline{D}} \underline{\underline{v}}^3 - \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{v}}^2 - \underline{\underline{v}}^{2*} \underline{\underline{D}} \underline{\underline{v}} \right) \\
&= \frac{3-a_1}{4} \underline{\underline{1}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} + \frac{1+a_1}{4} \left(\underline{\underline{h}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{v}} + \underline{\underline{v}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{h}} \right) \\
&\quad + \frac{a_1+3a_2-2}{8g} \left(\underline{\underline{1}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{v}}^3 - \underline{\underline{v}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{v}}^2 - \underline{\underline{v}}^{2T} \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{v}} \right) \\
&\stackrel{\text{SBP}}{=} \frac{3-a_1}{4} \underline{\underline{1}}^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} \underline{\underline{h}} \underline{\underline{v}} + \frac{1+a_1}{4} \underline{\underline{h}}^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} \underline{\underline{v}} \\
&\quad + \frac{a_1+3a_2-2}{8g} \left(\underline{\underline{1}}^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} \underline{\underline{v}}^3 - \underline{\underline{v}}^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} \underline{\underline{v}}^2 \right).
\end{aligned} \tag{86}$$

Here, $\underline{\underline{h}} \underline{\underline{1}} = \underline{\underline{h}}$, the SBP property $\underline{\underline{M}} \underline{\underline{D}} = \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} - \underline{\underline{D}}^T \underline{\underline{M}}$, and $\underline{\underline{D}} \underline{\underline{1}} = 0$ have been used. If multiplication and restriction to the boundary commute, these volume terms are simply $\underline{\underline{1}}^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} \underline{\underline{h}} \underline{\underline{v}}$ and yield the desired integral form. Similarly,

$$\begin{aligned}
\underline{\underline{1}}^T \underline{\underline{M}} \text{VOL}_{hv}^{a_1, a_2} &= \frac{1+a_1}{4} g \underline{\underline{1}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{h}}^2 + \frac{1-a_1}{2} g \underline{\underline{1}}^T \underline{\underline{M}} \underline{\underline{h}}^* \underline{\underline{D}} \underline{\underline{h}} - \frac{2a_1+3a_2-5}{8} \underline{\underline{1}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}}^2 \\
&\quad + \frac{2a_1+3a_2-1}{8} \underline{\underline{1}}^T \underline{\underline{M}} \left(\underline{\underline{h}}^* \underline{\underline{D}} \underline{\underline{v}}^2 + \underline{\underline{v}}^{2*} \underline{\underline{D}} \underline{\underline{h}} \right) + \frac{1}{2} \underline{\underline{1}}^T \underline{\underline{M}} \left(\underline{\underline{h}} \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{v}} + \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} \right) \\
&\quad + \frac{a_1+3a_2-2}{16g} \underline{\underline{1}}^T \underline{\underline{M}} \left(\underline{\underline{D}} \underline{\underline{v}}^4 - 2 \underline{\underline{v}}^{2*} \underline{\underline{D}} \underline{\underline{v}}^2 \right) + \frac{3-a_1}{4} g \underline{\underline{1}}^T \underline{\underline{M}} \underline{\underline{h}}^* \underline{\underline{D}} \underline{\underline{b}} \\
&\quad + \frac{1+a_1}{4} g \underline{\underline{1}}^T \underline{\underline{M}} \left(\underline{\underline{D}} \underline{\underline{h}} \underline{\underline{b}} - \underline{\underline{b}}^* \underline{\underline{D}} \underline{\underline{h}} \right) \\
&\quad + \frac{a_1+3a_2-2}{8} \underline{\underline{1}}^T \underline{\underline{M}} \left(\underline{\underline{D}} \underline{\underline{b}} \underline{\underline{v}}^2 - \underline{\underline{b}}^* \underline{\underline{D}} \underline{\underline{v}}^2 - 2 \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{b}} \underline{\underline{v}} + \underline{\underline{v}}^{2*} \underline{\underline{D}} \underline{\underline{b}} + 2 \underline{\underline{b}} \underline{\underline{v}}^* \underline{\underline{D}} \underline{\underline{v}} \right) \\
&= \frac{1+a_1}{4} g \underline{\underline{1}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{h}}^2 + \frac{1-a_1}{2} g \underline{\underline{h}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{h}} - \frac{2a_1+3a_2-5}{8} \underline{\underline{1}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}}^2 \\
&\quad + \frac{2a_1+3a_2-1}{8} \left(\underline{\underline{h}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{v}}^2 + \underline{\underline{v}}^{2T} \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{h}} \right) + \frac{1}{2} \left(\underline{\underline{h}} \underline{\underline{v}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{v}} + \underline{\underline{v}}^T \underline{\underline{M}} \underline{\underline{D}} \underline{\underline{h}} \underline{\underline{v}} \right)
\end{aligned} \tag{87}$$

$$\begin{aligned}
& + \frac{a_1 + 3a_2 - 2}{16g} \left(\underline{1}^T \underline{M} \underline{D} v^4 - 2v^{2T} \underline{M} \underline{D} v^2 \right) + \frac{3 - a_1}{4} gh^T \underline{M} \underline{D} b \\
& + \frac{1 + a_1}{4} g \left(\underline{1}^T \underline{M} \underline{D} \underline{h} b - \underline{b}^T \underline{M} \underline{D} \underline{h} \right) + \frac{a_1 + 3a_2 - 2}{8} \left(\underline{1}^T \underline{M} \underline{D} \underline{b} v^2 - \underline{b}^T \underline{M} \underline{D} v^2 \right) \\
& - \frac{a_1 + 3a_2 - 2}{8} \left(2v^T \underline{M} \underline{D} \underline{b} v - v^{2T} \underline{M} \underline{D} \underline{b} - 2bv^T \underline{M} \underline{D} v \right) \\
\stackrel{\text{SBP}}{=} & \frac{1 + a_1}{4} g \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{h}^2 + \frac{1 - a_1}{4} gh^T \underline{R}^T \underline{B} \underline{R} \underline{h} - \frac{2a_1 + 3a_2 - 5}{8} \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{h} v^2 \\
& + \frac{2a_1 + 3a_2 - 1}{8} \underline{h}^T \underline{R}^T \underline{B} \underline{R} v^2 + \frac{1}{2} v^T \underline{R}^T \underline{B} \underline{R} \underline{h} v + \frac{a_1 + 3a_2 - 2}{16g} \left[\underline{1}^T \underline{R}^T \underline{B} \underline{R} v^4 - v^{2T} \underline{R}^T \underline{B} \underline{R} v^2 \right] \\
& + \left\{ gh^T \underline{M} \underline{D} b \right\} - \frac{a_1 + 3a_2 - 2}{4} \left\{ v^T \underline{M} \underline{D} \underline{b} v - v^{2T} \underline{M} \underline{D} \underline{b} - \underline{b} v^T \underline{M} \underline{D} v \right\} \\
& + \frac{1 + a_1}{4} g \left[\underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{h} b - \underline{b}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \right] + \frac{a_1 + 3a_2 - 2}{8} \left[\underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{b} v^2 - \underline{b}^T \underline{R}^T \underline{B} \underline{R} v^2 \right].
\end{aligned}$$

Here, the terms in squared brackets $[\cdot]$ vanish if restriction to the boundary and multiplication commute, i.e. for a basis using Lobatto nodes. However, for other bases using e.g. Gauß nodes, these contributions are not zero in general. The first term in curly brackets $\{\cdot\}$ is a consistent discretisation of the source term $\int gh \partial_x b$. The second term in curly brackets $\{\cdot\}$ vanishes, if the product rule is valid, e.g. for constant bottom topography b . However, for general bottom topography, it is not of the desired form for the source influence $\int gh \partial_x b$ and it might be better to set it to zero by choosing $a_2 = \frac{2-a_1}{3}$, i.e. only the one-parameter family instead of the two-parameter family.

These surface terms obtained in (86) and (87) have to be balanced by the surface terms of the SBP SAT semidiscretisation (84) in order to get the desired result

$$\begin{aligned}
\underline{1}^T \underline{M} \partial_t \underline{h} &= - \underline{1}^T \underline{R}^T \underline{B} f_{\underline{h}}^{b_1, b_2}, \\
\underline{1}^T \underline{M} \partial_t \underline{h} v &= - \underline{1}^T \underline{R}^T \underline{B} f_{\underline{h} v}^{b_1, b_2} + \text{consistent contribution of } - \int gh \partial_x b,
\end{aligned} \tag{88}$$

leading to a conservative scheme.

Turning to *stability*, the approximation of

$$\int \partial_t U = \int U'(u) \cdot \partial_t u = \int w \cdot \partial_t u, \quad w = \left(g(h + b) - \frac{1}{2} v^2, v \right)^T, \tag{89}$$

influenced by the volume terms is given by

$$\begin{aligned}
& \underline{w}_1^T \underline{M} \text{VOL}_h^{a_1, a_2} + \underline{w}_2^T \underline{M} \text{VOL}_{\underline{h} v}^{a_1, a_2} \\
= & \frac{3 - a_1}{4} g (\underline{h} + \underline{b})^T \underline{M} \underline{D} \underline{h} v + \frac{1 + a_1}{4} g (\underline{h} + \underline{b})^T \underline{M} \left(\underline{h}^* \underline{D} v + \underline{v}^* \underline{D} \underline{h} \right) \\
& + \frac{a_1 + 3a_2 - 2}{8} (\underline{h} + \underline{b})^T \underline{M} \left(\underline{D} v^3 - \underline{v}^* \underline{D} v^2 - \underline{v}^{2*} \underline{D} v \right) - \frac{3 - a_1}{8} v^{2T} \underline{M} \underline{D} \underline{h} v \\
& - \frac{1 + a_1}{8} v^{2T} \underline{M} \left(\underline{h}^* \underline{D} v + \underline{v}^* \underline{D} \underline{h} \right) - \frac{a_1 + 3a_2 - 2}{16g} v^{2T} \underline{M} \left(\underline{D} v^3 - \underline{v}^* \underline{D} v^2 - \underline{v}^{2*} \underline{D} v \right) \\
& + \frac{1 + a_1}{4} g v^T \underline{M} \underline{D} \underline{h}^2 + \frac{1 - a_1}{2} g v^T \underline{M} \underline{h}^* \underline{D} \underline{h} - \frac{2a_1 + 3a_2 - 5}{8} v^T \underline{M} \underline{D} \underline{h} v^2 \\
& + \frac{2a_1 + 3a_2 - 1}{8} v^T \underline{M} \left(\underline{h}^* \underline{D} v^2 + \underline{v}^{2*} \underline{D} \underline{h} \right) + \frac{1}{2} v^T \underline{M} \left(\underline{h} v^* \underline{D} v + \underline{v}^* \underline{D} \underline{h} v \right) \\
& + \frac{a_1 + 3a_2 - 2}{16g} v^T \underline{M} \left(\underline{D} v^4 - 2v^{2*} \underline{D} v^2 \right) + \frac{3 - a_1}{4} g v^T \underline{M} \underline{h}^* \underline{D} \underline{b} \\
& + \frac{1 + a_1}{4} g v^T \underline{M} \left(\underline{D} \underline{h} b - \underline{b}^* \underline{D} \underline{h} \right) \\
& + \frac{a_1 + 3a_2 - 2}{8} v^T \underline{M} \left(\underline{D} \underline{b} v^2 - \underline{b}^* \underline{D} v^2 - 2v^* \underline{D} \underline{b} v + \underline{v}^{2*} \underline{D} \underline{b} + 2\underline{b} v^* \underline{D} v \right) \\
= & \frac{3 - a_1}{4} g (\underline{h} + \underline{b})^T \underline{M} \underline{D} \underline{h} v + \frac{1 + a_1}{4} g \left(\underline{h}^2 + \underline{b} \underline{h} \right)^T \underline{M} \underline{D} v + \frac{1 + a_1}{4} g (\underline{h} v + \underline{b} v)^T \underline{M} \underline{D} \underline{h} \\
& + \frac{a_1 + 3a_2 - 2}{8} (\underline{h} + \underline{b})^T \underline{M} \underline{D} v^3 - \frac{a_1 + 3a_2 - 2}{8} (\underline{h} v + \underline{b} v)^T \underline{M} \underline{D} v^2
\end{aligned} \tag{90}$$

$$\begin{aligned}
& -\frac{a_1 + 3a_2 - 2}{8} \left(hv^2 + bv^2 \right)^T \underline{\underline{M}} \underline{\underline{D}} v - \frac{3 - a_1}{8} v^{2T} \underline{\underline{M}} \underline{\underline{D}} hv - \frac{1 + a_1}{8} hv^{2T} \underline{\underline{M}} \underline{\underline{D}} v \\
& -\frac{1 + a_1}{8} v^{3T} \underline{\underline{M}} \underline{\underline{D}} h - \frac{a_1 + 3a_2 - 2}{16g} v^{2T} \underline{\underline{M}} \underline{\underline{D}} v^3 + \frac{a_1 + 3a_2 - 2}{16g} v^{3T} \underline{\underline{M}} \underline{\underline{D}} v^2 \\
& + \frac{a_1 + 3a_2 - 2}{16g} v^{4T} \underline{\underline{M}} \underline{\underline{D}} v + \frac{1 + a_1}{4} gv^T \underline{\underline{M}} \underline{\underline{D}} h^2 + \frac{1 - a_1}{2} ghv^T \underline{\underline{M}} \underline{\underline{D}} h \\
& -\frac{2a_1 + 3a_2 - 5}{8} v^T \underline{\underline{M}} \underline{\underline{D}} hv^2 + \frac{2a_1 + 3a_2 - 1}{8} hv^T \underline{\underline{M}} \underline{\underline{D}} v^2 + \frac{2a_1 + 3a_2 - 1}{8} v^{3T} \underline{\underline{M}} \underline{\underline{D}} h \\
& + \frac{1}{2} hv^{2T} \underline{\underline{M}} \underline{\underline{D}} v + \frac{1}{2} v^{2T} \underline{\underline{M}} \underline{\underline{D}} hv + \frac{a_1 + 3a_2 - 2}{16g} v^T \underline{\underline{M}} \underline{\underline{D}} v^4 \\
& -\frac{a_1 + 3a_2 - 2}{8g} v^{3T} \underline{\underline{M}} \underline{\underline{D}} v^2 + \frac{3 - a_1}{4} ghv^T \underline{\underline{M}} \underline{\underline{D}} b + \frac{1 + a_1}{4} gv^T \underline{\underline{M}} \underline{\underline{D}} hb \\
& -\frac{1 + a_1}{4} gbv^T \underline{\underline{M}} \underline{\underline{D}} h + \frac{a_1 + 3a_2 - 2}{8} v^T \underline{\underline{M}} \underline{\underline{D}} bv^2 - \frac{a_1 + 3a_2 - 2}{8} bv^T \underline{\underline{M}} \underline{\underline{D}} v^2 \\
& -\frac{a_1 + 3a_2 - 2}{4} v^{2T} \underline{\underline{M}} \underline{\underline{D}} bv + \frac{a_1 + 3a_2 - 2}{8} v^{3T} \underline{\underline{M}} \underline{\underline{D}} b \\
& + \frac{a_1 + 3a_2 - 2}{4} bv^{2T} \underline{\underline{M}} \underline{\underline{D}} v \\
& = \frac{3 - a_1}{4} g \left(h^T \underline{\underline{M}} \underline{\underline{D}} hv + hv^T \underline{\underline{M}} \underline{\underline{D}} h \right) + \frac{3 - a_1}{4} g \left(b^T \underline{\underline{M}} \underline{\underline{D}} hv + hv^T \underline{\underline{M}} \underline{\underline{D}} b \right) \\
& + \frac{1 + a_1}{4} g \left(h^{2T} \underline{\underline{M}} \underline{\underline{D}} v + v^T \underline{\underline{M}} \underline{\underline{D}} h^2 \right) + \frac{1 + a_1}{4} g \left(bh^T \underline{\underline{M}} \underline{\underline{D}} v + v^T \underline{\underline{M}} \underline{\underline{D}} bh \right) \\
& + 0 bv^T \underline{\underline{M}} \underline{\underline{D}} h + \frac{a_1 + 3a_2 - 2}{8} \left(h^T \underline{\underline{M}} \underline{\underline{D}} v^3 + v^{3T} \underline{\underline{M}} \underline{\underline{D}} h \right) + \frac{a_1 + 1}{8} \left(hv^T \underline{\underline{M}} \underline{\underline{D}} v^2 + v^{2T} \underline{\underline{M}} \underline{\underline{D}} hv \right) \\
& + \frac{5 - 2a_1 - 3a_2}{8} \left(hv^{2T} \underline{\underline{M}} \underline{\underline{D}} v + v^T \underline{\underline{M}} \underline{\underline{D}} hv^2 \right) - \frac{a_1 + 3a_2 - 2}{16g} \left(v^{2T} \underline{\underline{M}} \underline{\underline{D}} v^3 + v^{3T} \underline{\underline{M}} \underline{\underline{D}} v^2 \right) \\
& + \frac{a_1 + 3a_2 - 2}{16g} \left(v^{4T} \underline{\underline{M}} \underline{\underline{D}} v + v^T \underline{\underline{M}} \underline{\underline{D}} v^4 \right) + \frac{a_1 + 3a_2 - 2}{8} \left(bv^{2T} \underline{\underline{M}} \underline{\underline{D}} v + v^T \underline{\underline{M}} \underline{\underline{D}} bv^2 \right) \\
& -\frac{a_1 + 3a_2 - 2}{4} \left(bv^T \underline{\underline{M}} \underline{\underline{D}} v^2 + v^{2T} \underline{\underline{M}} \underline{\underline{D}} bv \right) + \frac{a_1 + 3a_2 - 2}{8} \left(b^T \underline{\underline{M}} \underline{\underline{D}} v^3 + v^{3T} \underline{\underline{M}} \underline{\underline{D}} b \right) \\
& \stackrel{\text{SBP}}{=} \frac{3 - a_1}{4} gh^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} hv + \frac{3 - a_1}{4} gb^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} hv + \frac{1 + a_1}{4} gv^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h^2 \\
& + \frac{1 + a_1}{4} gv^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} bh + \frac{a_1 + 3a_2 - 2}{8} h^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^3 + \frac{a_1 + 1}{8} hv^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^2 \\
& + \frac{5 - 2a_1 - 3a_2}{8} v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} hv^2 - \frac{a_1 + 3a_2 - 2}{16g} v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^3 + \frac{a_1 + 3a_2 - 2}{16g} v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^4 \\
& + \frac{a_1 + 3a_2 - 2}{8} v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} bv^2 - \frac{a_1 + 3a_2 - 2}{4} bv^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^2 + \frac{a_1 + 3a_2 - 2}{8} b^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^3.
\end{aligned}$$

If multiplication and restriction to the boundary commute, these terms simplify to $\underline{\underline{1}}^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} \underline{\underline{F}}$, where $\underline{\underline{F}} = gh^2v + gbhv + \frac{1}{2}hv^3$ is the entropy flux (14).

These surface contributions resulting from the volume terms have to be balanced by the surface terms $w_1^T \underline{\underline{M}} \text{SURF}_h + w_2^T \underline{\underline{M}} \text{SURF}_{hv}$, in order to get an estimate of the form $\llbracket w \rrbracket \cdot f^{\text{num}} - \llbracket \psi \rrbracket$ for the entropy change influenced by one boundary node (if the bottom topography is continuous across elements). That is, the simple interpolations

$$\underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} hv, \quad \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} hv^2, \quad \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h^2, \quad (91)$$

in the surface terms (83) for the method including boundary nodes have to be adapted.

The following combination of surface term structures proposed by Ortleb (2016); Ranocha, Öffner, and Sonar (2016) will be investigated

$$\begin{aligned}
& \text{SURF}_h^{a_1, a_2} \\
& = b_1 \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} hv + b_2 \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} h)(\underline{\underline{R}} v) + b_3 h^* \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v + b_4 v^* \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h \\
& + \frac{c_1}{g} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^3 + \frac{c_2}{g} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v)(\underline{\underline{R}} v^2) + \frac{c_3}{g} v^* \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^2 \\
& + \frac{c_4}{g} v^{2*} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v,
\end{aligned} \quad (92)$$

$$\begin{aligned}
& \underline{\text{SURF}}_{hv}^{a_1, a_2} \\
&= d_1 \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v}^2 + d_2 \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{h}) (\underline{R} \underline{v}^2) + d_3 \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{h} \underline{v}) (\underline{R} \underline{v}) \\
&+ d_4 \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{h}) (\underline{R} \underline{v})^2 + d_5 \underline{v}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v} + d_6 \underline{h} \underline{v}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{v} \\
&+ d_7 \underline{v}^{2*} \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{h} + d_8 \underline{h}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{v}^2 \\
&+ e_1 g \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{h}^2 + e_2 g \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{h})^2 + e_3 g \underline{h}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{h} \\
&+ \frac{k_1}{g} \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{v}^4 + \frac{k_2}{g} \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{v}) (\underline{R} \underline{v}^3) + \frac{k_3}{g} \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{v}^2)^2 \\
&+ \frac{k_4}{g} \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{v})^2 (\underline{R} \underline{v}^2) + \frac{k_5}{g} \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{v})^4 + \frac{k_6}{g} \underline{v}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{v}^3 \\
&+ \frac{k_7}{g} \underline{v}^* \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{v}) (\underline{R} \underline{v}^2) + \frac{k_8}{g} \underline{v}^* \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{v})^3 + \frac{k_9}{g} \underline{v}^{2*} \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{v}^2 \\
&+ \frac{k_{10}}{g} \underline{v}^{2*} \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{v})^2 + \frac{k_{11}}{g} \underline{v}^{3*} \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{v} \\
&+ l_1 \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{b} \underline{v}^2 + l_2 \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{b}) (\underline{R} \underline{v}^2) + l_3 \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{b} \underline{v}) (\underline{R} \underline{v}) \\
&+ l_4 \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{b}) (\underline{R} \underline{v})^2 + l_5 \underline{b}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{v}^2 + l_6 \underline{v}^{2*} \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{b} \\
&+ l_7 \underline{b} \underline{v}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{v} + l_8 \underline{v}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{b} \underline{v} \\
&+ l_9 \underline{b}^* \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{v})^2 + l_{10} \underline{v}^* \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{b}) (\underline{R} \underline{v}) \\
&+ m_1 g \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{b} \underline{h} + m_2 g \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{b}) (\underline{R} \underline{h}) \\
&+ m_3 g \underline{h}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{b} + m_4 g \underline{b}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{h} \\
&- \frac{1}{2} \underline{v}^* \underline{M}^{-1} \underline{R}^T \underline{B} \underline{f}_h^{a_1, a_2} + \frac{1}{2} \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{f}_h^{a_1, a_2}) (\underline{R} \underline{v}),
\end{aligned}$$

where $b_i, c_i, d_i, e_i, k_i, l_i, m_i \in \mathbb{R}$ are free parameters that have to be determined.

Considering *conservation* for h , the relevant conditions are obtained by multiplying the surface terms with $\underline{1}^T \underline{M}$.

$$\begin{aligned}
& \underline{1}^T \underline{M} \underline{\text{SURF}}_h^{a_1, a_2} \\
&= b_1 \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v} + b_2 \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{h}) (\underline{R} \underline{v}) + b_3 \underline{h}^T \underline{R}^T \underline{B} \underline{R} \underline{v} + b_4 \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \\
&+ \frac{c_1}{g} \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^3 + \frac{c_2}{g} \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{v}) (\underline{R} \underline{v}^2) + \frac{c_3}{g} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^2 + \frac{c_4}{g} \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{v} \\
&= b_1 \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v} + (b_2 + b_3 + b_4) \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{h}) (\underline{R} \underline{v}) \\
&+ c_1 \frac{1}{g} \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^3 + (c_2 + c_3 + c_4) \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{v}) (\underline{R} \underline{v}^2).
\end{aligned} \tag{93}$$

Here, some manipulations of the form

$$\underline{h}^T \underline{R}^T \underline{B} \underline{R} \underline{v} = \underline{h}_R \underline{v}_R - \underline{h}_L \underline{v}_L = \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{h}) (\underline{R} \underline{v}) \tag{94}$$

as proposed by Ranocha, Öffner, and Sonar (2016) have been used. Thus, comparison with (86) yields the conditions

$$\begin{aligned}
b_1 &= \frac{3 - a_1}{4}, & b_2 + b_3 + b_4 &= \frac{1 + a_1}{4}, \\
c_1 &= \frac{a_1 + 3a_2 - 2}{8}, & c_2 + c_3 + c_4 &= -\frac{a_1 + 3a_2 - 2}{8}.
\end{aligned} \tag{95}$$

Similarly, for hv ,

$$\begin{aligned}
& \underline{1}^T \underline{M} \underline{\text{SURF}}_{hv}^{a_1, a_2} \\
&= d_1 \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v}^2 + d_2 \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{h}) (\underline{R} \underline{v}^2) + d_3 \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{h} \underline{v}) (\underline{R} \underline{v}) + d_4 \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{h}) (\underline{R} \underline{v})^2 \\
&+ d_5 \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v} + d_6 \underline{h} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{v} + d_7 \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{h} + d_8 \underline{h}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^2
\end{aligned} \tag{96}$$

$$\begin{aligned}
& + e_1 g \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{h}^2 + e_2 g \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{h})^2 + e_3 g \underline{h}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \\
& + \frac{k_1}{g} \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^4 + \frac{k_2}{g} \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})(\underline{R} \underline{v}^3) + \frac{k_3}{g} \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{v}^2)^2 \\
& + \frac{k_4}{g} \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})^2 (\underline{R} \underline{v}^2) + \frac{k_5}{g} \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})^4 + \frac{k_6}{g} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^3 \\
& + \frac{k_7}{g} \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})(\underline{R} \underline{v}^2) + \frac{k_8}{g} \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})^3 + \frac{k_9}{g} \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{v}^2 \\
& + \frac{k_{10}}{g} \underline{v}^{2T} \underline{M}^{-1} \underline{R}^T \underline{B} (\underline{R} \underline{v})^2 + \frac{k_{11}}{g} \underline{v}^{3T} \underline{M}^{-1} \underline{R}^T \underline{B} \underline{R} \underline{v} \\
& + l_1 \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{b} \underline{v}^2 + l_2 \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{b})(\underline{R} \underline{v}^2) + l_3 \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{b} \underline{v})(\underline{R} \underline{v}) + l_4 \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{b})(\underline{R} \underline{v})^2 \\
& + l_5 \underline{b}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^2 + l_6 \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{b} + l_7 \underline{b} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{v} \\
& + l_8 \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{b} \underline{v} + l_9 \underline{b}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})^2 + l_{10} \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{b})(\underline{R} \underline{v}) \\
& + m_1 g \underline{1}^T \underline{R}^T \underline{B} \underline{R} \underline{b} \underline{h} + m_2 g \underline{1}^T \underline{R}^T \underline{B} (\underline{R} \underline{b})(\underline{R} \underline{h}) + m_3 g \underline{h}^T \underline{R}^T \underline{B} \underline{R} \underline{b} + m_4 g \underline{b}^T \underline{R}^T \underline{B} \underline{R} \underline{h}.
\end{aligned}$$

Analogously, comparing this with (87) results in the conditions

$$\begin{aligned}
d_1 &= \frac{5 - 2a_1 - 3a_2}{8}, & d_2 + d_7 + d_8 &= \frac{2a_1 + 3a_2 - 1}{8}, & (97) \\
d_3 + d_5 + d_6 &= \frac{1}{2}, & d_4 &= 0, \\
e_1 &= \frac{1 + a_1}{4}, & e_2 + e_3 &= \frac{1 - a_1}{4}, \\
k_1 &= \frac{a_1 + 3a_2 - 2}{16}, & k_2 + k_6 + k_{11} &= 0, \\
k_3 + k_9 &= -\frac{a_1 + 3a_2 - 2}{16}, & k_4 + k_7 + k_{10} &= 0, \\
k_5 + k_8 &= 0, & l_1 &= \frac{a_1 + 3a_2 - 2}{8}, \\
l_2 + l_5 + l_6 &= -\frac{a_1 + 3a_2 - 2}{8}, & l_3 + l_7 + l_8 &= 0, \\
l_4 + l_9 + l_{10} &= 0, & m_1 &= \frac{1 + a_1}{4}, \\
m_2 + m_3 + m_4 &= -\frac{1 + a_1}{4}.
\end{aligned}$$

Considering *stability*, the surface terms (92) yield

$$\begin{aligned}
& \underline{w}_1^T \underline{M} \text{SURF}_h^{a_1, a_2} + \underline{w}_2^T \underline{M} \text{SURF}_{hv}^{a_1, a_2} \\
& = b_1 g \underline{h}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v} + b_2 g \underline{h}^T \underline{R}^T \underline{B} (\underline{R} \underline{h})(\underline{R} \underline{v}) + b_3 g \underline{h}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{v} + b_4 g \underline{h} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \\
& + c_1 \underline{h}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^3 + c_2 \underline{h}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})(\underline{R} \underline{v}^2) + c_3 \underline{h} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^2 + c_4 \underline{h} \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{v} \\
& + b_1 g \underline{b}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v} + b_2 g \underline{b}^T \underline{R}^T \underline{B} (\underline{R} \underline{h})(\underline{R} \underline{v}) + b_3 g \underline{b} \underline{h}^T \underline{R}^T \underline{B} \underline{R} \underline{v} + b_4 g \underline{b} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \\
& + c_1 \underline{b}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^3 + c_2 \underline{b}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})(\underline{R} \underline{v}^2) + c_3 \underline{b} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^2 + c_4 \underline{b} \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{v} \\
& - \frac{1}{2} b_1 \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v} - \frac{1}{2} b_2 \underline{v}^{2T} \underline{R}^T \underline{B} (\underline{R} \underline{h})(\underline{R} \underline{v}) - \frac{1}{2} b_3 \underline{h} \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{v} - \frac{1}{2} b_4 \underline{v}^{3T} \underline{R}^T \underline{B} \underline{R} \underline{h} \\
& - \frac{c_1}{2g} \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{v}^3 - \frac{c_2}{2g} \underline{v}^{2T} \underline{R}^T \underline{B} (\underline{R} \underline{v})(\underline{R} \underline{v}^2) - \frac{c_3}{2g} \underline{v}^{3T} \underline{R}^T \underline{B} \underline{R} \underline{v}^2 - \frac{c_4}{2g} \underline{v}^{4T} \underline{R}^T \underline{B} \underline{R} \underline{v} \\
& + d_1 \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v}^2 + d_2 \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{h})(\underline{R} \underline{v}^2) + d_3 \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{h} \underline{v})(\underline{R} \underline{v}) + d_4 \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{h})(\underline{R} \underline{v})^2 \\
& + d_5 \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{h} \underline{v} + d_6 \underline{h} \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{v} + d_7 \underline{v}^{3T} \underline{R}^T \underline{B} \underline{R} \underline{h} + d_8 \underline{h} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^2 \\
& + e_1 g \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{h}^2 + e_2 g \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{h})^2 + e_3 g \underline{h} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{h} \\
& + \frac{k_1}{g} \underline{v}^T \underline{R}^T \underline{B} \underline{R} \underline{v}^4 + \frac{k_2}{g} \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})(\underline{R} \underline{v}^3) + \frac{k_3}{g} \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{v}^2)^2 \\
& + \frac{k_4}{g} \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})^2 (\underline{R} \underline{v}^2) + \frac{k_5}{g} \underline{v}^T \underline{R}^T \underline{B} (\underline{R} \underline{v})^4 + \frac{k_6}{g} \underline{v}^{2T} \underline{R}^T \underline{B} \underline{R} \underline{v}^3
\end{aligned}
\tag{98}$$

$$\begin{aligned}
& + \frac{k_7}{g} v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v) (\underline{\underline{R}} v^2) + \frac{k_8}{g} v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v)^3 + \frac{k_9}{g} v^{3T} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^2 \\
& + \frac{k_{10}}{g} v^{3T} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v)^2 + \frac{k_{11}}{g} v^{4T} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v \\
& + l_1 v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} \underline{\underline{B}} v^2 + l_2 v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} b) (\underline{\underline{R}} v^2) + l_3 v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} b v) (\underline{\underline{R}} v) + l_4 v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} b) (\underline{\underline{R}} v)^2 \\
& + l_5 b v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^2 + l_6 v^{3T} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} b + l_7 b v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v \\
& + l_8 v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} b v + l_9 b v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v)^2 + l_{10} v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} b) (\underline{\underline{R}} v) \\
& + m_1 g v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} b h + m_2 g v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} b) (\underline{\underline{R}} h) + m_3 g h v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} b + m_4 g b v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h \\
& - \frac{1}{2} v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} f_h^{a_1, a_2} + \frac{1}{2} v^T \underline{\underline{R}}^T \underline{\underline{B}} (f_h^{a_1, a_2}) (\underline{\underline{R}} v) \\
= & (b_1 + b_4 + e_3) g h^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h v + (b_2) g h^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} h) (\underline{\underline{R}} v) + (b_3 + e_1) g h^{2T} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v \\
& + \left(c_1 - \frac{1}{2} b_4 + d_7 \right) h^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^3 + \left(c_2 - \frac{1}{2} b_2 + d_2 \right) h^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v) (\underline{\underline{R}} v^2) \\
& + \left(c_3 - \frac{1}{2} b_1 + d_5 \right) h v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^2 + \left(c_4 - \frac{1}{2} b_3 + d_1 + d_6 \right) h v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v \\
& + (b_1 + m_3) g b^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h v + (b_2 + m_2) g b^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} h) (\underline{\underline{R}} v) + (b_3 + m_1) g b h^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v \\
& + (b_4 + m_4) g b v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h + (c_1 + l_6) b^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^3 + (c_2 + l_2 + l_{10}) b^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v) (\underline{\underline{R}} v^2) \\
& + (c_3 + l_5 + l_8) b v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^2 + (c_4 + l_1 + l_7) b v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v \\
& + \left(-\frac{1}{2} c_1 - \frac{1}{2} c_3 + k_6 + k_9 \right) \frac{1}{g} v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^3 + \left(-\frac{1}{2} c_2 + k_3 + k_7 \right) \frac{1}{g} v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v) (\underline{\underline{R}} v^2) \\
& + \left(-\frac{1}{2} c_4 + k_1 + k_{11} \right) \frac{1}{g} v^{4T} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v + d_3 v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} h v) (\underline{\underline{R}} v) + d_4 v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} h) (\underline{\underline{R}} v)^2 \\
& + d_8 h v^T \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} v^2 + e_2 g v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} h)^2 \\
& + (k_2 + k_{10}) \frac{1}{g} v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v) (\underline{\underline{R}} v^3) + (k_4 + k_8) \frac{1}{g} v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v)^2 (\underline{\underline{R}} v^2) \\
& + k_5 \frac{1}{g} v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} v)^4 + (l_3 + l_9) v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} b v) (\underline{\underline{R}} v) + l_4 v^T \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} b) (\underline{\underline{R}} v)^2 \\
& - \frac{1}{2} v^{2T} \underline{\underline{R}}^T \underline{\underline{B}} f_h^{a_1, a_2} + \frac{1}{2} v^T \underline{\underline{R}}^T \underline{\underline{B}} (f_h^{a_1, a_2}) (\underline{\underline{R}} v).
\end{aligned}$$

Comparing this with (90) yields the conditions

$$\begin{aligned}
b_1 + b_4 + e_3 &= \frac{3 - a_1}{4}, & b_2 + e_2 &= \frac{1}{2}, \\
b_3 + e_1 &= \frac{1 + a_1}{4}, & c_1 - \frac{b_4}{2} + d_7 &= \frac{a_1 + 3a_2 - 2}{8}, \\
c_2 - \frac{b_2}{2} + d_2 &= 0, & c_3 - \frac{b_1}{2} + d_5 + d_8 &= \frac{a_1 + 1}{8}, \\
c_4 - \frac{b_3}{2} + d_1 + d_6 &= \frac{5 - 2a_1 - 3a_2}{8}, & b_1 + m_3 &= \frac{3 - a_1}{4}, \\
b_2 + m_2 &= 0, & b_3 + m_1 &= \frac{a_1 + 1}{4}, \\
b_4 + m_4 &= 0, & c_1 + l_6 &= \frac{a_1 + 3a_2 - 2}{8}, \\
c_2 + l_2 + l_{10} &= 0, & c_3 + l_5 + l_8 &= -\frac{a_1 + 3a_2 - 2}{4}, \\
c_4 + l_1 + l_7 &= \frac{a_1 + 3a_2 - 2}{8}, & l_4 &= 0, \\
-\frac{c_1}{2} - \frac{c_3}{2} + k_6 + k_9 &= -\frac{a_1 + 3a_2 - 2}{16}, & -\frac{c_2}{2} + k_3 + k_7 &= 0, \\
-\frac{c_4}{2} + k_1 + k_{11} &= \frac{a_1 + 3a_2 - 2}{16}, & d_3 &= 0, \\
d_4 &= 0, & k_2 + k_{10} &= 0,
\end{aligned} \tag{99}$$

$$\begin{aligned}k_4 + k_8 &= 0, \\ l_3 + l_9 &= 0,\end{aligned}$$

$$\begin{aligned}k_5 &= 0 \\ l_4 &= 0.\end{aligned}$$

Solving the linear system given by (95), (97), and (99) with SymPy (SymPy Development Team, 2016) results in the free parameters $m_4, k_9, k_{10}, k_{11}, l_{10}$ for any given parameters a_1, a_2 :

$$\begin{aligned}b_1 &= -\frac{a_1}{4} + \frac{3}{4}, & b_2 &= \frac{a_1}{4} + m_4 + \frac{1}{4}, \\ b_3 &= 0, & b_4 &= -m_4, \\ c_1 &= \frac{a_1}{8} + \frac{3a_2}{8} - \frac{1}{4}, & c_2 &= -\frac{a_1}{8} - \frac{3a_2}{8} - 2k_{10} - 2k_9 + \frac{1}{4}, \\ c_3 &= 2k_{10} - 2k_{11} + 2k_9, & c_4 &= 2k_{11}, \\ d_1 &= -\frac{a_1}{4} - \frac{3a_2}{8} + \frac{5}{8}, & d_2 &= \frac{2a_1 + 3a_2 - 1}{8} + 2k_{10} + 2k_9 + \frac{m_4}{2}, \\ d_3 &= 0, & d_4 &= 0, \\ d_5 &= 2k_{11} + \frac{1}{2}, & d_6 &= -2k_{11}, \\ d_7 &= -\frac{m_4}{2}, & d_8 &= -2k_{10} - 2k_9, \\ e_1 &= \frac{a_1}{4} + \frac{1}{4}, & e_2 &= -\frac{a_1}{4} - m_4 + \frac{1}{4}, \\ e_3 &= m_4, & k_1 &= \frac{a_1}{16} + \frac{3a_2}{16} - \frac{1}{8}, \\ k_2 &= -k_{10}, & k_3 &= -\frac{a_1}{16} - \frac{3a_2}{16} - k_9 + \frac{1}{8}, \\ k_4 &= 0, & k_5 &= 0, \\ k_6 &= k_{10} - k_{11}, & k_7 &= -k_{10}, \\ k_8 &= 0, & l_1 &= \frac{a_1}{8} + \frac{3a_2}{8} - \frac{1}{4}, \\ l_2 &= \frac{a_1 + 3a_2 - 2}{8} + 2k_{10} + 2k_9 - l_{10}, & l_3 &= l_{10}, \\ l_4 &= 0, & l_5 &= l_{10} - \frac{a_1 + 3a_2 - 2}{4} - 2k_{10} - 2k_9, \\ l_6 &= 0, & l_7 &= -2k_{11}, \\ l_8 &= 2k_{11} - l_{10}, & l_9 &= -l_{10}, \\ m_1 &= \frac{a_1}{4} + \frac{1}{4}, & m_2 &= -\frac{a_1}{4} - m_4 - \frac{1}{4}, \\ m_3 &= 0.\end{aligned} \tag{100}$$

This proves the following

Theorem 8. For $a_1, a_2, \alpha_1, \alpha_2 \in \mathbb{R}$, using a general SBP operator, the semidiscretisation

$$\begin{aligned}\partial_t \underline{h} &= -\underline{\text{VOL}}_h^{a_1, a_2} + \underline{\text{SURE}}_h^{a_1, a_2} - \underline{M}^{-1} \underline{R}^T \underline{B} f_h^{\alpha_1, \alpha_2}, \\ \partial_t \underline{h} \underline{v} &= -\underline{\text{VOL}}_{hv}^{a_1, a_2} + \underline{\text{SURE}}_{hv}^{a_1, a_2} - \underline{M}^{-1} \underline{R}^T \underline{B} f_{hv}^{\alpha_1, \alpha_2},\end{aligned} \tag{101}$$

with volume terms (85) and surface terms (92), where the parameters are chosen according to (100) with free parameters $m_4, k_9, k_{10}, k_{11}, l_{10} \in \mathbb{R}$,

1. conserves the total mass $\int h$. Additionally, it conserves the total momentum $\int hv$, if the bottom topography is constant. Otherwise, the rate of change is consistent with the source term $-gh\partial_x b$.
2. conserves the total entropy / energy $\int U$.
3. handles the lake-at-rest condition correctly.

That is, this semidiscretisation is conservative (across elements), stable (entropy conservative), and well-balanced.

For an SBP operator using a nodal basis with diagonal mass matrix $\underline{\underline{M}}$, the volume terms can be equivalently expressed in the flux difference form corresponding to the extended numerical fluxes (80) in Lemma 6.

For the special case $a_2 = \frac{2-a_1}{3}$, i.e. the one-parameter family, the volume terms (85) can be considerably simplified. Accordingly, the ansatz (92) can be simplified by setting the coefficients c_i , k_i , and l_i to zero. In this case, only m_4 remains as free parameter. In this case, the surface terms (92) become

$$\begin{aligned} \text{SURE}_h^{a_1, \frac{2-a_1}{3}} &= \frac{3-a_1}{4} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h v + \frac{a_1+1+4m_4}{4} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} h) (\underline{\underline{R}} v) \\ &\quad - m_4 \underline{\underline{v}}^* \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h, \\ \text{SURE}_{hv}^{a_1, \frac{2-a_1}{3}} &= \frac{3-a_1}{8} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h v^2 + \frac{a_1+1+4m_4}{8} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} h) (\underline{\underline{R}} v^2) \\ &\quad + \frac{1}{2} \underline{\underline{v}}^* \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h v - \frac{m_4}{2} \underline{\underline{v}}^{2*} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h + \frac{a_1+1}{4} g \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h^2 \\ &\quad + \frac{1-a_1-4m_4}{4} g \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} h)^2 + m_4 g h^* \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h \\ &\quad + \frac{a_1+1}{4} g \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} b h - \frac{a_1+1+4m_4}{4} g \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} (\underline{\underline{R}} b) (\underline{\underline{R}} h) \\ &\quad + m_4 g b^* \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} \underline{\underline{R}} h - \frac{1}{2} \underline{\underline{v}}^* \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} f_h^{a_1, a_2} + \frac{1}{2} \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} (f_h^{a_1, a_2}) (\underline{\underline{R}} v). \end{aligned} \tag{102}$$

Here, the choice $m_4 = 0$ results in the fewest terms. Additionally, choosing $a_1 = -1$ cancels some other terms and results in the skew-symmetric form of Gassner, Winters, and Kopriva (2016a).

The two-parameter family of fluxes (48) has been derived in entropy variables w and translated to primitive variables h, v (49) in the following calculations. Hence, it may seem natural to consider a splitting similar to (85) and (92) but expressed using entropy instead of primitive variables. The volume terms can be obtained similarly to (85), but the surface terms are more delicate. A general ansatz similar to (92) without the correction $-\underline{\underline{v}}^* \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} f_h^{a_1, a_2} + \underline{\underline{M}}^{-1} \underline{\underline{R}}^T \underline{\underline{B}} (f_h^{a_1, a_2}) (\underline{\underline{R}} v)$ has not been successful, i.e. the resulting linear system was not solvable. However, it might be possible to get the desired result using another ansatz for the surface terms.

7 Positivity preservation

In this section, the framework of positivity preservation by Zhang and Shu (2011) will be presented and adapted to the setting of a nodal SBP method. The argumentation of Zhang and Shu (2011) can be summarised as

1. Since the method should be conservative, ensure a non-negative mean value of h in each cell.
2. If the height becomes negative somewhere but the mean value is non-negative, use a suitable limiter to enforce the non-negativity as needed.

Considering the semidiscretisation of the water height h , the discrete total mass in the standard element evolves as

$$\underline{\underline{1}}^T \underline{\underline{M}} \partial_t \underline{\underline{h}} = -\underline{\underline{1}}^T \underline{\underline{R}}^T \underline{\underline{B}} f_h^{\text{num}}, \tag{103}$$

since the method is conservative. Therefore, the mean value in a cell i of width Δx evolves as

$$\partial_t \bar{h}_i = -\frac{1}{\Delta x} (f_R^{\text{num}} - f_L^{\text{num}}), \tag{104}$$

where $f_{R,L}^{\text{num}}$ is the numerical flux between cells at the left / right hand side. Using a forward Euler method as time discretisation,

$$\bar{h}_i^+ = \bar{h}_i - \frac{\Delta t}{\Delta x} (f_R^{\text{num}} - f_L^{\text{num}}). \quad (105)$$

Here, the numerical fluxes depend on the value of the variables at the boundaries. However, in a pure finite volume framework, in cell i

$$\bar{h}_i^+ = \bar{h}_i - \frac{\Delta t}{\Delta x} (f^{\text{num}}(\bar{u}_i, \bar{u}_{i+1}) - f^{\text{num}}(\bar{u}_{i-1}, \bar{u}_i)). \quad (106)$$

If there are numerical fluxes such that under a suitable CFL condition $\Delta t \leq c\Delta x$ the non-negativity can be guaranteed, this can be extended to the DG setting. Choose $q+1$ Lobatto-Legendre nodes x_k in the element i . Since Lobatto quadrature with weight ω_k at x_k is exact for polynomials of degree $2q-1$, ensure $2q-1 \geq p$, i.e. $q \geq \frac{p-1}{2}$. Then, $\bar{h}_i = \sum_{k=0}^q \omega_k h_{i,k}$, where $h_{i,k} = h_i(x_k)$. Since boundary nodes are included,

$$\begin{aligned} \bar{h}_i^+ = & \omega_0 \left(h_{i,0} - \frac{\Delta t}{\omega_0 \Delta x} (f^{\text{num}}(u_{i,0}, u_{i,q}) - f^{\text{num}}(u_{i-1,q}, u_{i,0})) \right) + \sum_{k=1}^{q-1} \omega_k h_{i,k} \\ & + \omega_q \left(h_{i,q} - \frac{\Delta t}{\omega_q \Delta x} (f^{\text{num}}(u_{i,q}, u_{i+1,0}) - f^{\text{num}}(u_{i,0}, u_{i,q})) \right), \end{aligned} \quad (107)$$

where $\omega_0 = \omega_q$ has been used to add zero. Thus, if the finite volume method (106) using the numerical flux f^{num} is positivity preserving under the CFL condition $\Delta t \leq c\Delta x$, the DG method is positivity preserving under the scaled CFL condition $\Delta t \leq \omega_q c\Delta x$, if the values of $h_{i,k}$ at the quadrature nodes x_0, \dots, x_1 are non-negative.

Then, after the Euler step, the mean value in each cell i is ensured to be non-negative. Applying the simple linear scaling limiter

$$h_i \mapsto \tilde{h}_i := \theta h_i + (1 - \theta) \bar{h}_i = \bar{h}_i + \theta (h_i - \bar{h}_i), \quad \theta = \begin{cases} 1, & \min h_i \geq 0, \\ \frac{\bar{h}_i}{h_i - \min h_i}, & \min h_i < 0, \end{cases} \quad (108)$$

of Liu and Osher (1996, Section 2.3) ensures $\min h_i \geq 0$. Here, $\min h_i$ can be computed as the minimum of the polynomial h_i of degree $\leq p$ in the complete cell, or as the minimum of h_i at the nodes x_0, \dots, x_q used to guarantee the non-negativity of the cell mean. Applying this limiter in each Euler sub-step, this procedure can be extended to SSP methods consisting of convex combinations of Euler steps.

In a nodal collocation framework such as nodal DG or FD, it would be more natural to enforce the non-negativity of the water height at the collocation nodes ξ_0, \dots, ξ_p in the standard element. Thus, if Lobatto nodes are used, the framework of Zhang and Shu (2011) described above extends directly, where the non-negativity of the mean value \bar{h}_i^+ can be guaranteed under the possibly worse CFL condition $\Delta t \leq \omega_p c\Delta x$, where $p \geq q$, when the limiter (108) is applied with $\min h_i = \min \{h_i(\xi_0), \dots, h_i(\xi_p)\}$. Alternatively, to get a better CFL condition $\Delta t \leq \omega_q c\Delta x$, the water height can be interpolated to the other nodes x_0, \dots, x_q and the limiter (108) can be applied with $\min h_i = \min \{h_i(x_0), \dots, h_i(x_q), h_i(\xi_0), \dots, h_i(\xi_p)\}$.

Similarly, if Gauß nodes are considered, only one possibility is obvious: Interpolate to suitable Lobatto nodes x_0, \dots, x_q and use the limiter (108), again with the choice of the minimum as minimum over solution nodes and interpolation points $\min h_i = \min \{h_i(x_0), \dots, h_i(x_q), h_i(\xi_0), \dots, h_i(\xi_p)\}$.

If this limiter should be coupled with an entropy conservative / stable method, a natural question is whether the limiter (108) is entropy stable. This is indeed true, since for a convex entropy U

$$\begin{aligned} \overline{U(\tilde{h}_i)} &= \overline{\theta U(h_i) + (1 - \theta) U(\bar{h}_i)} \\ &\stackrel{U \text{ convex}}{\leq} \theta \overline{U(h_i)} + (1 - \theta) \overline{U(\bar{h}_i)} = \theta \overline{U(h_i)} + (1 - \theta) U(\bar{h}_i) \\ &\stackrel{\text{Jensen}}{\leq} \theta \overline{U(h_i)} + (1 - \theta) \overline{U(\bar{h}_i)} = \overline{U(h_i)}, \end{aligned} \quad (109)$$

where the monotonicity of the mean value, the convexity of U , and Jensen's inequality have been used. This proves

Lemma 9. *The scaling limiter (108) is entropy stable for $\theta \in [0, 1]$.*

8 Numerical surface fluxes / Riemann solvers

In this section, several numerical surface fluxes will be presented, that can be used to get entropy stable, positivity preserving, and well-balanced schemes.

Often, intensive research involving numerical fluxes / Riemann solvers is done in a discrete finite volume setting, where the update procedure can be written as

$$u_i^+ = u_i - \frac{\Delta t}{\Delta x} (f^{\text{num}}(u_{i+1}, u_i) - f^{\text{num}}(u_i, u_{i-1})). \quad (110)$$

In such a fully discrete setting, a natural discrete entropy inequality is

$$U(u_i^+) \leq U(u_i) - \frac{\Delta t}{\Delta x} (F^{\text{num}}(u_{i+1}, u_i) - F^{\text{num}}(u_i, u_{i-1})), \quad (111)$$

obtained under a suitable CFL condition $\Delta t \leq c\Delta x$, where F^{num} is a consistent numerical entropy flux. As described by Bouchut (2004, Section 2.2.2), this implies in the semi-discrete limit

$$\begin{aligned} F(u_R) + w(u_R) \cdot (f^{\text{num}}(u_L, u_R) - f(u_R)) &\leq F^{\text{num}}(u_L, u_R), \text{ and} \\ F^{\text{num}}(u_L, u_R) &\leq F(u_L) + w(u_L) \cdot (f^{\text{num}}(u_L, u_R) - f(u_L)). \end{aligned} \quad (112)$$

Inserting the flux potential $\psi = w \cdot f - F$, this condition can be restated as

$$w(u_R) \cdot f^{\text{num}}(u_L, u_R) - \psi(u_R) \leq F^{\text{num}}(u_L, u_R) \leq w(u_L) \cdot f^{\text{num}}(u_L, u_R) - \psi(u_L). \quad (113)$$

If the left hand side is smaller than the right hand side, any numerical flux F^{num} between these two values will be an adequate numerical entropy flux, especially the arithmetic mean value of both values, the numerical entropy flux chosen by Tadmor (1987)

$$\begin{aligned} F^{\text{num}}(u_L, u_R) &= \underbrace{\frac{w(u_L) + w(u_R)}{2}}_{= \{w\}} \cdot f^{\text{num}}(u_L, u_R) - \underbrace{\frac{\psi(u_L) + \psi(u_R)}{2}}_{= \{\psi\}}. \end{aligned} \quad (114)$$

This condition (113) can be restated as the condition for an entropy stable numerical flux given by Tadmor (1987)

$$\underbrace{(w(u_R) - w(u_L))}_{= [w]} \cdot f^{\text{num}}(u_L, u_R) - \underbrace{(\psi(u_R) - \psi(u_L))}_{= [\psi]} \leq 0. \quad (115)$$

Thus, if a numerical flux fulfils a fully discrete entropy condition (111), it does also fulfil the semidiscrete entropy stability condition (115).

8.1 Entropy conservative fluxes for constant bottom topography b

Here, the one-parameter entropy conservative flux (36)

$$f_h^\alpha = \frac{1 - \alpha}{2} \{hv\} + \frac{1 + \alpha}{2} \{h\} \{v\} \quad (116)$$

will be considered. Inserting this in a FV evolution equation (106) (and dropping the bars $\bar{\cdot}$),

$$\begin{aligned} h_i^+ &= h_i - \frac{\Delta t}{\Delta x} (f_h^\alpha(u_i, u_{i+1}) - f_h^\alpha(u_{i-1}, u_i)) \\ &= h_i - \frac{\Delta t}{\Delta x} \left(\frac{1 - \alpha}{2} \frac{h_i v_i + h_{i+1} v_{i+1}}{2} + \frac{1 + \alpha}{2} \frac{h_i + h_{i+1}}{2} \frac{v_i + v_{i+1}}{2} \right) \end{aligned} \quad (117)$$

$$\begin{aligned}
& -\frac{1-\alpha}{2} \frac{h_i v_i + h_{i-1} v_{i-1}}{2} - \frac{1+\alpha}{2} \frac{h_i + h_{i-1}}{2} \frac{v_i + v_{i-1}}{2} \Big) \\
& = h_i \left[1 - \frac{\Delta t}{\Delta x} \frac{1+\alpha}{8} (v_{i+1} - v_{i-1}) \right] - h_{i+1} \frac{\Delta t}{\Delta x} \left[\frac{3-\alpha}{8} v_{i+1} + \frac{1+\alpha}{8} v_i \right] \\
& \quad + h_{i-1} \frac{\Delta t}{\Delta x} \left[\frac{3-\alpha}{8} v_{i-1} + \frac{1+\alpha}{8} v_i \right].
\end{aligned}$$

Considering non-negative water height h , if $h_i = 0$, $h_{i-1}, h_{i+1} > 0$, $v_i = 0$, $(3-\alpha)v_{i-1} < 0$, $(3-\alpha)v_{i+1} > 0$, the height h_i^+ becomes negative for $\Delta t > 0$.

If only positive water height h should be considered, using the same conditions as before but $h_i > 0$, the new water height can be guaranteed to be positive, but only under a CFL condition depending on h_i with allowed $\Delta t \rightarrow 0$ as $h_i \rightarrow 0$.

Lemma 10. *The one-parameter family of entropy conservative fluxes (36) for the shallow water equations with constant bottom topography b is not positivity preserving under a CFL condition not blowing up as $h \rightarrow 0$.*

If the full two-parameter family (49) of entropy conservative fluxes is considered, terms of the form $v_{i\pm 1}^3 - v_{i\pm 1}^2 v_i - v_{i\pm 1} v_i^2 + v_i^3$ have to be added. Since these terms may be of arbitrary size and sign and do not contain any multiple of the water height h , they can render the water height negative. Thus, this family is not positivity preserving, too.

8.2 Adding dissipation for constant bottom topography b

Classically, adding "dissipation" to some kind of central flux f^{cent} would result in

$$f^{\text{cent}} - P \llbracket u \rrbracket, \quad (118)$$

where P is positive semi-definite. However, this might not result in an entropy stable scheme. Therefore, the amount of dissipation is better chosen as proportional to the jump of entropy variables $\llbracket w \rrbracket$ instead of $\llbracket u \rrbracket$. In the simplest case, $P = \lambda I$ with $\lambda \geq 0$. Then,

$$\lambda \llbracket u \rrbracket \approx \lambda \frac{\partial u}{\partial w} \llbracket w \rrbracket. \quad (119)$$

Since $\partial_w u = (\partial_u w)^{-1}$, $\partial_w u = U''(u)$, and U is convex, this dissipation matrix is positive semi-definite. Thus, the resulting numerical flux

$$f^{\text{num}} = f^\alpha - \frac{\lambda}{2} \overline{\partial_w u} \llbracket w \rrbracket \quad (120)$$

is entropy stable, where f^α is the one-parameter entropy conservative flux (36) and $\overline{\partial_w u}$ is a suitable positive semi-definite approximation of the entropy Jacobian $\partial_w u$, e.g. the Jacobian evaluated at some mean value \bar{u} .

For the shallow water equations (1) (cf. (4) and (17)),

$$\partial_u w = \begin{pmatrix} g + \frac{v^2}{h} & -\frac{v}{h} \\ -\frac{v}{h} & \frac{1}{h} \end{pmatrix}, \quad \partial_w u = \begin{pmatrix} \frac{1}{g} & \frac{v}{g} \\ \frac{v}{g} & h + \frac{v^2}{g} \end{pmatrix}. \quad (121)$$

Thus, adding dissipation in the form (120), the following additional contribution has to be added to the right hand side of (117) for the entropy variables $w = \left(gh - \frac{1}{2}v^2, v \right)^T$, i.e. if the bottom topography b is continuous across cell boundaries

$$\begin{aligned}
& \frac{\Delta t}{2\Delta x} \left[\lambda_{i+\frac{1}{2}} \overline{\partial_w u}_{i+\frac{1}{2}} (w_{i+1} - w_i) - \lambda_{i-\frac{1}{2}} \overline{\partial_w u}_{i-\frac{1}{2}} (w_i - w_{i-1}) \right]_h \\
& = \frac{\Delta t}{\Delta x} \frac{\lambda_{i+\frac{1}{2}}}{2} \left(h_{i+1} - \frac{v_{i+1}^2}{2g} - h_i + \frac{v_i^2}{2g} + \frac{\bar{v}_{i+\frac{1}{2}}}{g} (v_{i+1} - v_i) \right)
\end{aligned} \quad (122)$$

$$+ \frac{\Delta t}{\Delta x} \frac{\lambda_{i-\frac{1}{2}}}{2} \left(-h_i + \frac{v_i^2}{2g} + h_{i-1} - \frac{v_{i-1}^2}{2g} - \frac{\bar{v}_{i-\frac{1}{2}}}{g} (v_i - v_{i-1}) \right).$$

The additional positive values of $h_{i\pm 1}$ are crucial to obtain the positivity preserving property under a suitable CFL condition, if λ is large enough. The negative values of h_i are weighted by Δt and can thus be bounded by the positive terms in (117) if Δt is small enough. The remaining terms containing only the velocity but not the height of the water may be problematic. However, if the Jacobian $\overline{\partial_u w}$ is evaluated at the arithmetic mean value,

$$\bar{v}_{i+\frac{1}{2}} = \llbracket v \rrbracket_{i,i+1} = \frac{v_{i+1} + v_i}{2} \Rightarrow \bar{v}_{i+\frac{1}{2}}(v_{i+1} - v_i) = \frac{1}{2}v_{i+1}^2 - \frac{1}{2}v_i^2, \quad (123)$$

and similarly for $\bar{v}_{i-\frac{1}{2}}$. Thus, these additional terms vanish and (cf. equation (117))

$$\begin{aligned} h_i^+ = & h_i \left[1 - \frac{\Delta t}{\Delta x} \frac{1+\alpha}{8} (v_{i+1} - v_{i-1}) - \frac{\lambda_{i+\frac{1}{2}} + \lambda_{i-\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x} \right] \\ & + h_{i+1} \frac{\Delta t}{\Delta x} \left[\frac{\lambda_{i+\frac{1}{2}}}{2} - \frac{3-\alpha}{8} v_{i+1} - \frac{1+\alpha}{8} v_i \right] + h_{i-1} \frac{\Delta t}{\Delta x} \left[\frac{\lambda_{i-\frac{1}{2}}}{2} + \frac{3-\alpha}{8} v_{i-1} + \frac{1+\alpha}{8} v_i \right]. \end{aligned} \quad (124)$$

Hence, for $\lambda_{i\pm\frac{1}{2}} \geq \max\{|v_i|, |v_{i\pm 1}|\}$, the coefficients of $h_{i\pm 1}$ are non-negative, and under the CFL condition

$$\frac{\Delta t}{\Delta x} \left(\left| \frac{1+\alpha}{8} \right| (|v_{i+1}| + |v_{i-1}|) + \frac{\lambda_{i+\frac{1}{2}} + \lambda_{i-\frac{1}{2}}}{2} \right) \leq 1, \quad (125)$$

the new height h_i^+ is non-negative. Note that this CFL condition does not blow up as $h_i \rightarrow 0$. With this estimate, the choice $\alpha = -1$ seems to be optimal in order to get the least restrictive CFL condition. Again, considering the full two-parameter family (49) instead of the one-parameter family (37) results in additional terms of order v^3 not containing any contribution of the water height h . Thus, these terms are of arbitrary size and sign and destroy the positivity preservation as in

Lemma 11. *The one-parameter local Lax-Friedrichs type flux (120) is entropy stable, if $\lambda \overline{\partial_u w}$ is positive semi-definite. Additionally, it is positivity preserving under the CFL condition (125), if*

$$\overline{\partial_u w} = \partial_u w (\llbracket h \rrbracket, \llbracket v \rrbracket), \quad \lambda \geq \max\{|v_-|, |v_+|\}, \quad (126)$$

and the bottom topography b is continuous across cell boundaries.

In the implementation, $\lambda = \max\{|v_-| + \sqrt{gh_-}, |v_+| + \sqrt{gh_+}\}$ is chosen, as in the classical Lax-Friedrichs flux.

However, if the bottom topography b is discontinuous across cell boundaries, additional terms have to be considered, since the entropy variables are $w = \left(g(h+b) - \frac{1}{2}v^2, v\right)^T$. These terms are

$$\frac{\Delta t}{\Delta x} \frac{\lambda_{i+\frac{1}{2}}}{2} (b_{i+1} - b_i) + \frac{\Delta t}{\Delta x} \frac{\lambda_{i-\frac{1}{2}}}{2} (-b_i + b_{i-1}). \quad (127)$$

Adding these terms to the right hand side of equation (124), the CFL condition (125) gets lost. If all heights are positive, it is possible to guarantee $h_i^+ \geq 0$, but the corresponding CFL condition blows up as $h \rightarrow 0$, since $b_{i\pm 1} - b_i$ may be of arbitrary size and has to be balanced by positive contributions of the h terms.

With the choice of $\overline{\partial_u w}$ as in Lemma 11, the dissipation term becomes

$$\begin{aligned} & - \frac{\lambda}{2} \partial_u w (\llbracket u \rrbracket) [\llbracket w \rrbracket] \\ & = - \frac{\lambda}{2} \begin{pmatrix} \frac{1}{g} & \frac{\llbracket v \rrbracket}{g} \\ \frac{\llbracket v \rrbracket}{g} & \llbracket h \rrbracket + \frac{\llbracket v \rrbracket^2}{g} \end{pmatrix} \begin{pmatrix} g [\llbracket h + b \rrbracket] - \frac{1}{2} [\llbracket v^2 \rrbracket] \\ [\llbracket v \rrbracket] \end{pmatrix} \end{aligned} \quad (128)$$

$$\begin{aligned}
&= -\frac{\lambda}{2} \left(\begin{aligned} &[h+b] - \frac{\{v\}}{g} [v] + \frac{\{v\}}{g} [v] \\ &\{v\} [h+b] - \frac{\{v\}^2}{g} [v] + \{h\} [v] + \frac{\{v\}^2}{g} [v] \end{aligned} \right) \\
&= -\frac{\lambda}{2} \left(\begin{aligned} &[h+b] \\ &[hv] + \{v\} [b] \end{aligned} \right),
\end{aligned}$$

where the product rule (30) has been used. Thus, if b is continuous across cell boundaries, $[b] = 0$ and the dissipation term is simply the classical local Lax-Friedrichs dissipation term $-\frac{\lambda}{2} [u]$.

The same numerical flux can also be obtained as Rusanov type flux. Choosing a dissipation approximately as

$$-|f'(u)| [u] \approx -\left| \frac{\partial f}{\partial u} \right| \frac{\partial u}{\partial w} [w] \quad (129)$$

and using the scaling of Barth (1999, Theorem 4) for the eigenvectors, resulting in

$$\left| \frac{\partial f}{\partial u} \right| = R|\Lambda| R^{-1}, \quad \frac{\partial u}{\partial w} = R R^T, \quad (130)$$

where Λ is the diagonal matrix of eigenvalues of $f'(u)$, yields

$$-|f'(u)| [u] \approx -R|\Lambda| R^T [w]. \quad (131)$$

Thus, the Rusanov choice of dissipation with $|\Lambda| = \lambda I$, where $\lambda > 0$ is the largest eigenvalue, yields exactly the same Lax Friedrichs type dissipative flux (120).

A Roe type dissipation operator can be constructed by choosing

$$|\Lambda| = \text{diag}(|\lambda|_i), \quad (132)$$

where λ_i are the eigenvalues of $f'(u)$. Computing the Roe type dissipation $-R|\Lambda| R^T [w]$ with $|\Lambda| = \text{diag}(|\lambda|_-, |\lambda|_+)$ results for a continuous bottom topography b using the scaled eigenvectors (18) and the product rule (30) in

$$\begin{aligned}
&-R|\Lambda| R^T [w] \\
&= -\frac{1}{2g} \begin{pmatrix} 1 & 1 \\ \lambda_- & \lambda_+ \end{pmatrix} \begin{pmatrix} |\lambda_-| & 0 \\ 0 & |\lambda_+| \end{pmatrix} \begin{pmatrix} 1 & \lambda_- \\ 1 & \lambda_+ \end{pmatrix} \begin{pmatrix} g[h] - \frac{1}{2}[v^2] \\ [v] \end{pmatrix} \\
&= -\frac{1}{2g} \begin{pmatrix} |\lambda_-| & |\lambda_+| \\ \lambda_-|\lambda_-| & \lambda_+|\lambda_+| \end{pmatrix} \begin{pmatrix} g[h] - \{v\}[v] + v[v] - \sqrt{gh}[v] \\ g[h] - \{v\}[v] + v[v] + \sqrt{gh}[v] \end{pmatrix} \\
&= -\frac{1}{2} \left((|\lambda_-| + |\lambda_+|) [h] + \frac{|\lambda_-| + |\lambda_+|}{g} (v - \{v\}) [v] - \frac{|\lambda_-| + |\lambda_+|}{g} \sqrt{gh} [v] \right).
\end{aligned} \quad (133)$$

However, some attempts to prove the preservation of positivity have not been successful.

8.3 Some classical numerical fluxes for constant bottom topography b

Using a Godunov scheme with exact solution of the Riemann problem results in an entropy stable and positivity preserving scheme, since the exact solution has these properties. However, some nonlinear root finding algorithm has to be applied to compute the exact solution of the Riemann problem. Therefore, this will not be considered here in more detail. The Riemann problem for the shallow water equations with vanishing bottom topography is described in detail inter alia by Holden and Risebro (2002, Chapter 5).

Since some details of the solution of the Riemann problem are lumped by taking the average, Harten, Lax, and Leer (1983) proposed an approximate Riemann solver using only one intermediate state, known as HLL Riemann solver, using estimates of the slowest and fastest wave speeds s_- , s_+ . If these estimates are lower and upper bounds of the wave speeds in the solution of the Riemann problem, this flux is entropy stable, as remarked by Harten, Lax, and Leer

(1983). Additionally, by the right choice of wave speed estimates as in the famous HLLE version for Euler equations proposed by Einfeldt (1988), the numerical flux is positivity preserving, similar to the results for gas dynamics established by Einfeldt, Munz, Roe, and Sjögreen (1991), since the intermediate state in the approximating solution of the Riemann problem satisfies this condition.

Not only the local Lax Friedrichs type numerical flux (120) in the previous section is positivity preserving and entropy stable, but also its classical variant

$$f^{\text{num}} = \llbracket f \rrbracket - \frac{\lambda}{2} \llbracket u \rrbracket. \quad (134)$$

This can be established using general results of Bouchut (2003); Frid (2001, 2004), but also by direct calculation.

Using the FV update procedure (106), the water height after one time step using the local Lax Friedrichs flux (134) becomes

$$\begin{aligned} h_i^+ &= h_i - \frac{\Delta t}{\Delta x} (f^{\text{num}}(u_i, u_{i+1}) - f^{\text{num}}(u_{i-1}, u_i)) \\ &= h_i - \frac{\Delta t}{\Delta x} \left(\frac{h_{i+1}v_{i+1} + h_i v_i}{2} - \frac{\lambda_{i+\frac{1}{2}}}{2} (h_{i+1} - h_i) - \frac{h_i v_i + h_{i-1}v_{i-1}}{2} + \frac{\lambda_{i-\frac{1}{2}}}{2} (h_i - h_{i-1}) \right) \\ &= h_i \left[1 - \frac{\Delta t}{\Delta x} \frac{\lambda_{i-\frac{1}{2}} + \lambda_{i+\frac{1}{2}}}{2} \right] + h_{i+1} \frac{\lambda_{i+\frac{1}{2}} - v_{i+1}}{2} + h_{i-1} \frac{\lambda_{i-\frac{1}{2}} - v_{i-1}}{2}. \end{aligned} \quad (135)$$

Thus, if $\lambda_{i\pm\frac{1}{2}} > v_{i\pm 1}$ and $\Delta t < \frac{2\Delta x}{\lambda_{i-\frac{1}{2}} + \lambda_{i+\frac{1}{2}}}$, the new water height h_i^+ is non-negative, if the previous water heights h_{i-1}, h_i, h_{i+1} are non-negative.

The entropy stability in the semidiscrete setting with vanishing bottom topography b can be established by

$$\begin{aligned} &\llbracket w \rrbracket \cdot f^{\text{num}} - \llbracket \psi \rrbracket \\ &= \left(g \llbracket h \rrbracket - \frac{1}{2} \llbracket v^2 \rrbracket \right) \left(\llbracket hv \rrbracket - \frac{\lambda}{2} \llbracket h \rrbracket \right) + \llbracket v \rrbracket \left(\llbracket hv^2 \rrbracket + \frac{1}{2} g \llbracket h^2 \rrbracket - \frac{\lambda}{2} \llbracket hv \rrbracket \right) - \frac{1}{2} g \llbracket h^2 v \rrbracket \\ &= g \llbracket hv \rrbracket \llbracket h \rrbracket - \llbracket hv \rrbracket \llbracket v \rrbracket \llbracket v \rrbracket - \frac{\lambda}{2} g \llbracket h \rrbracket^2 + \frac{\lambda}{2} \llbracket v \rrbracket \llbracket h \rrbracket \llbracket v \rrbracket + \llbracket hv^2 \rrbracket \llbracket v \rrbracket \\ &\quad + \frac{1}{2} g \llbracket h^2 \rrbracket \llbracket v \rrbracket - \frac{\lambda}{2} \llbracket h \rrbracket \llbracket v \rrbracket^2 - \frac{\lambda}{2} \llbracket v \rrbracket \llbracket h \rrbracket \llbracket v \rrbracket - \frac{1}{2} g \llbracket h^2 v \rrbracket \\ &= -\frac{\lambda}{2} g \llbracket h \rrbracket^2 - \frac{\lambda}{2} \llbracket h \rrbracket \llbracket v \rrbracket^2 + g \llbracket hv \rrbracket \llbracket h \rrbracket - \llbracket hv \rrbracket \llbracket v \rrbracket \llbracket v \rrbracket + \llbracket hv^2 \rrbracket \llbracket v \rrbracket - g \llbracket h \rrbracket \llbracket v \rrbracket \llbracket h \rrbracket, \end{aligned} \quad (136)$$

where the product rule (30) has been used. Direct calculation yields

$$\begin{aligned} &\llbracket w \rrbracket \cdot f^{\text{num}} - \llbracket \psi \rrbracket \\ &= -\frac{1}{2} g \underbrace{(h_+ - h_-)^2}_{\geq 0} \underbrace{\left(\lambda - \frac{v_+ - v_-}{2} \right)}_{\geq 0} - \frac{1}{4} \underbrace{h_+}_{\geq 0} \underbrace{(\lambda - v_+)}_{\geq 0} \underbrace{(v_+ - v_-)^2}_{\geq 0} - \frac{1}{4} \underbrace{h_-}_{\geq 0} \underbrace{(\lambda - v_-)}_{\geq 0} \underbrace{(v_+ - v_-)^2}_{\geq 0} \\ &\geq 0, \end{aligned} \quad (137)$$

if $h_+, h_- \geq 0$ and $\lambda \geq |v_+|, |v_-|$. This proves

Lemma 12. *The local Lax-Friedrichs flux (134) used in a simple FV method for the shallow water equations with constant bottom topography b is positivity preserving, if the water height is non-negative and the CFL condition*

$$1 - \frac{\Delta t}{\Delta x} \frac{\lambda_{i-\frac{1}{2}} + \lambda_{i+\frac{1}{2}}}{2} \geq 0 \quad (138)$$

is fulfilled. Additionally, it is entropy stable in the semidiscrete sense, if the water heights are non-negative ($h_+, h_- \geq 0$) and $\lambda \geq |v_+|, |v_-|$.

In the implementation, $\lambda = \max \left\{ |v_-| + \sqrt{gh_-}, |v_+| + \sqrt{gh_+} \right\}$ is chosen.

8.4 Suliciu relaxation solver for constant bottom topography b

The Suliciu relaxation solver described by Bouchut (2004, Section 2.4) for the shallow water equations has been implemented with some technical modifications to allow vanishing water height h as follows.

At first, intermediate speeds are computed as

$$\begin{aligned} \text{if } h_+ \geq h_- \wedge h_+ > 0, \quad & \begin{cases} \frac{c_-}{h_-} = \sqrt{gh_-} + \frac{3}{2} \max \left\{ 0, \frac{g}{2} \frac{h_+^2 - h_-^2}{h_+ \sqrt{gh_+}} + v_- - v_+ \right\}, \\ \frac{c_+}{h_+} = \sqrt{gh_+} + \frac{3}{2} \max \left\{ 0, \frac{g}{2} \frac{h_-^2 - h_+^2}{c_-} + v_- - v_+ \right\}, \end{cases} \\ \text{if } h_+ \leq h_- \wedge h_- > 0, \quad & \begin{cases} \frac{c_+}{h_+} = \sqrt{gh_+} + \frac{3}{2} \max \left\{ 0, \frac{g}{2} \frac{h_-^2 - h_+^2}{h_- \sqrt{gh_-}} + v_- - v_+ \right\}, \\ \frac{c_-}{h_-} = \sqrt{gh_-} + \frac{3}{2} \max \left\{ 0, \frac{g}{2} \frac{h_+^2 - h_-^2}{c_+} + v_- - v_+ \right\}, \end{cases} \\ \text{if } h_- = h_+ = 0, \quad & \begin{cases} \frac{c_-}{h_-} = \frac{c_+}{h_+} = 0. \end{cases} \end{aligned} \quad (139)$$

Then, intermediate values are computed with $c_{\pm} = h_{\pm} \frac{c_{\pm}}{h_{\pm}}$ as

$$\begin{aligned} v_-^* = v_+^* &= \frac{c_- v_- + c_+ v_+ + \frac{g}{2}(h_-^2 - h_+^2)}{c_- + c_+}, & \pi_-^* = \pi_+^* &= \frac{\frac{g}{2}(c_+ h_-^2 + c_- h_+^2) - c_- c_+ (v_+ - v_-)}{c_- + c_+}, \\ h_-^* &= \left(\frac{1}{h_-} + \frac{c_+ (v_+ - v_-) + \frac{g}{2}(h_-^2 - h_+^2)}{c_- (c_- + c_+)} \right)^{-1}, & h_+^* &= \left(\frac{1}{h_+} + \frac{c_- (v_+ - v_-) + \frac{g}{2}(h_+^2 - h_-^2)}{c_+ (c_- + c_+)} \right)^{-1}, \end{aligned} \quad (140)$$

where the quantities $v_-^*, v_+^*, \pi_-^*, \pi_+^*, h_-^*, h_+^*$ have been set to zero if their numerical evaluation returned NaN (not a number, e.g. if a division "0/0" occurs). Finally, the numerical fluxes are given by

$$\begin{pmatrix} f_h^{\text{num}} \\ f_{hv}^{\text{num}} \end{pmatrix} (h_-, v_-, h_+, v_+) = \begin{cases} \begin{pmatrix} h_- v_- \\ h_- v_-^2 + \frac{g}{2} h_-^2 \end{pmatrix}, & \text{if } 0 \leq v_- - \frac{c_-}{h_-}, \\ \begin{pmatrix} h_-^* v_-^* \\ h_-^* v_-^{*2} + \pi_-^* \end{pmatrix}, & \text{if } v_- - \frac{c_-}{h_-} < 0 \leq v_-^* \equiv v_+^*, \\ \begin{pmatrix} h_+^* v_+^* \\ h_+^* v_+^{*2} + \pi_+^* \end{pmatrix}, & \text{if } v_-^* \equiv v_+^* < 0 \leq v_+ + \frac{c_+}{h_+}, \\ \begin{pmatrix} h_+ v_+ \\ h_+ v_+^2 + \frac{g}{2} h_+^2 \end{pmatrix}, & \text{else.} \end{cases} \quad (141)$$

This numerical flux is entropy stable and positivity preserving, where

$$\frac{\Delta t}{\Delta x} \max \left\{ \left| v_- - \frac{c_-}{h_-} \right|, \left| v_+ + \frac{c_+}{h_+} \right| \right\} \leq \frac{1}{2} \quad (142)$$

is the corresponding CFL condition for a fully discrete scheme, see Bouchut (2004, Chapter 2). However,

$$\frac{\Delta t}{\Delta x} \max \left\{ \left| v_- - \frac{c_-}{h_-} \right|, \left| v_+ + \frac{c_+}{h_+} \right| \right\} \leq 1 \quad (143)$$

suffices for positivity preservation and semidiscrete entropy stability, similarly to the CFL conditions for the local Lax-Friedrichs flux (138)

8.5 Kinetic solver for constant bottom topography b

Using a kinetic approach, Perthame and Simeoni (2001) proposed a fully discrete finite volume method for the shallow water equations with general bottom topography b that has the three desired properties, i.e. that is entropy stable, positivity preserving, and well-balanced under a suitable CFL condition not blowing up as $h \rightarrow 0$. As described at the beginning of this section, the semidiscrete entropy stability follows. However, the corresponding numerical flux has to be evaluated by some quadrature if the bottom topography varies. This will not be used here. But in the case of a constant topography, all integrals can be evaluated analytically, resulting in the following scheme.

If the left and right state is given by h_-, v_- and h_+, v_+ , respectively, the numerical fluxes are the integrals

$$\begin{pmatrix} f_h^{\text{num}} \\ f_{hv}^{\text{num}} \end{pmatrix} = \int_{\xi \geq 0} \begin{pmatrix} \xi \\ \xi^2 \end{pmatrix} M_-(\xi) d\xi + \int_{\xi \leq 0} \begin{pmatrix} \xi \\ \xi^2 \end{pmatrix} M_+(\xi) d\xi, \quad (144)$$

where M is the corresponding Maxwellian

$$M_{\pm}(\xi) = M(\xi, h_{\pm}, v_{\pm}) = \sqrt{\frac{2h_{\pm}}{g}} \frac{1}{\pi} \sqrt{\max \left\{ 0, 1 - \frac{(\xi - v_{\pm})^2}{2\pi g} \right\}}. \quad (145)$$

Using the symmetry of the kinetic integrals (by a substitution $\xi \rightarrow -\xi$)

$$\int_{\xi \leq 0} \begin{pmatrix} \xi \\ \xi^2 \end{pmatrix} M(\xi, h, v) d\xi = \int_{\xi \geq 0} \begin{pmatrix} -\xi \\ \xi^2 \end{pmatrix} M(\xi, h, -v) d\xi, \quad (146)$$

it suffices to compute the integrals $\int_{\xi \geq 0}$. Using Mathematica (Wolfram Research, Inc. 2014), these can be expressed after some additional simplifications as

$$\int_{\xi \geq 0} \xi M(\xi, h, v) d\xi = \begin{cases} hv, & \\ \text{if } h > 0 \wedge v \geq \sqrt{2gh}, & \\ \frac{1}{2}hv + \frac{2}{3\pi}h\sqrt{2gh-v^2} + \frac{1}{6g\pi}v^2\sqrt{2gh-v^2} + \frac{1}{\pi}hv \arctan \frac{v}{\sqrt{2gh-v^2}}, & \\ \text{if } h > 0 \wedge \sqrt{2gh} + v > 0 \wedge \sqrt{2gh} - v > 0, & \\ 0, & \\ \text{else,} & \end{cases} \quad (147)$$

and

$$\begin{aligned} & \int_{\xi \geq 0} \xi^2 M(\xi, h, v) d\xi \\ &= \begin{cases} hv^2 + \frac{1}{2}gh^2, & \\ \text{if } h > 0 \wedge v \geq \sqrt{2gh}, & \\ \frac{hv^2 + \frac{1}{2}gh^2}{2} + \frac{13}{12\pi}hv\sqrt{2gh-v^2} + \frac{1}{12g\pi}v^3\sqrt{2gh-v^2} + \frac{1}{\pi} \left(hv^2 + \frac{1}{2}gh^2 \right) \arctan \frac{v}{\sqrt{2gh-v^2}}, & \\ \text{if } h > 0 \wedge \sqrt{2gh} + v > 0 \wedge \sqrt{2gh} - v > 0, & \\ 0, & \\ \text{else.} & \end{cases} \end{aligned} \quad (148)$$

The corresponding CFL condition is

$$\frac{\Delta t}{\Delta x} \max \left\{ |v_-| + \sqrt{2gh_-}, |v_+| + \sqrt{2gh_+} \right\} \leq 1. \quad (149)$$

8.6 Hydrostatic reconstruction approach for general bottom topography b

The hydrostatic reconstruction has been introduced by Audusse, Bouchut, Bristeau, Klein, and Perthame (2004) as a means to extend a numerical flux for the shallow water equations with constant bottom topography b to varying b , preserving good properties of the numerical flux, notably entropy stability and positivity preservation. In addition, the resulting method is well-balanced for the lake-at-rest initial condition.

In the context of extended numerical fluxes incorporating the source term, the flux using hydrostatic reconstruction can be described as follows: Compute at first the limited values $\tilde{h}_i = \max \{0, h_i + b_i - \max \{b_i, b_k\}\}$, $\tilde{h}_k = \max \{0, h_k + b_k - \max \{b_i, b_k\}\}$ and use the extended numerical fluxes with new arguments

$$f_h^{\text{num}}(\tilde{h}_i, v_i, \tilde{h}_k, v_k) \quad (150)$$

for the water height h and

$$f_{hv}^{\text{num}}(\tilde{h}_i, v_i, \tilde{h}_k, v_k) + \frac{g}{2}(h_i^2 - \tilde{h}_i^2) \quad (151)$$

for the discharge hv to compute the rate of change in cell i influenced by cell k .

This results in a consistent and well-balanced numerical flux that is positivity preserving and entropy stable, if the given fluxes f_h^{num} , f_{hv}^{num} have these properties for the shallow water equations with constant bottom b .

However, this hydrostatic reconstruction has some disadvantages for some combinations of bottom slope, mesh size, and water height, as described by Delestre, Cordier, Darboux, and James (2012), at least if used for a first order FV scheme.

8.7 Other approaches for general bottom topography b

By an approach based on relaxation, Berthon and Chalons (2016) constructed an approximate Riemann solver that has the three desired properties, i.e. that is entropy stable, positivity preserving, and well-balanced under a suitable CFL condition. However, the existence of some parameters and a suitable time step are based on asymptotic arguments and can therefore not be implemented directly.

However, the shallow water equations are derived based on the assumption of low variations in the bottom topography. Hence, a discretisation of it that is continuous across elements seems to be natural.

9 Finite volume subcells

Although the analysis of the previous sections suggests that the semidiscretisation of Theorem 8 with appropriate positivity preserving and entropy stable fluxes of section 8 and the positivity preserving limiter of Zhang and Shu (2011), described also in section 7, is stable, there are problems at wet-dry fronts in the practical implementation. These problems can be handled by some appropriate limiting strategy, e.g. TVB limiters used by Xing, Zhang, and Shu (2010) or the slope limiter used by Duran and Marche (2014). However, since the high order of the approximation is lost in these cases, the approach of finite volume subcells used similarly by Meister and Ortleb (2016) in the context of the shallow water equations will be pursued. Additionally, given the interpretation of SBP methods with diagonal norm as subcell flux differencing methods by Fisher and Carpenter (2013); Fisher, Carpenter, Nordström, Yamaleev, and Swanson (2013), the usage of FV subcells seems to be quite natural.

In order to use finite volume subcells to compute the time derivative, the general procedure can be described as follows:

1. Decide, whether the high-order discretisation or FV subcells of first order should be used.
2. Project the polynomial of degree $\leq p$ (using $p + 1$ degrees of freedom) onto a piecewise constant solution with $p + 1$ subcells.

3. Compute the classical FV time derivative.

As a detector to use FV subcells, the water height in the element or adjacent elements will be used, as described in section 10.

The projection in step 2 is done for a diagonal-norm nodal SBP basis simply by taking subcells of length $M_{i,i}$ with corresponding value u_i . This is not an exact projection for the polynomial u in general, but is very simple and fits to the subcell flux differencing framework of Fisher and Carpenter (2013); Fisher, Carpenter, Nordström, Yamaleev, and Swanson (2013).

Thus, for an SBP SAT semidiscretisation

$$\partial_t u = -\underline{\text{VOL}} + \underline{\text{SURF}} - \underline{M}^{-1} \underline{R}^T \underline{B} f^{\text{num}}, \quad (152)$$

the surface terms $\underline{\text{SURF}}$ are set to zero and the numerical flux f^{num} is computed using the outer values u_0, u_p instead of a higher order interpolation $\underline{R}u$ – for nodal bases including boundary points, this makes no difference. The volume terms $\underline{\text{VOL}}$ are computed via FV subcells as

$$\begin{aligned} [\underline{\text{VOL}}]_0 &= \frac{f_{0,1}^{\text{num}}}{M_{0,0}}, \\ [\underline{\text{VOL}}]_i &= \frac{f_{i,i+1}^{\text{num}} - f_{i,i-1}^{\text{num}}}{M_{i,i}}, \quad i \in \{1, \dots, p-1\}, \\ [\underline{\text{VOL}}]_p &= -\frac{f_{p,p-1}^{\text{num}}}{M_{p,p}}. \end{aligned} \quad (153)$$

Therefore, the numerical flux terms $\underline{M}^{-1} \underline{R}^T \underline{B} f^{\text{num}}$ and the volume terms $\underline{\text{VOL}}$ form together a finite volume discretisation of the subcells.

10 Numerical tests

In this section, some numerical tests of the proposed schemes will be performed. For the time integration, the third-order, three stage SSP method SSPRK(3,3) given by Gottlieb and Shu (1998) will be used: $\underline{u} \mapsto \underline{u}^+$, and the stages are

$$\begin{aligned} \underline{u}_1 &= \underline{u} + \Delta t \partial_t \underline{u}, \\ \underline{u}_2 &= \frac{3}{4} \underline{u} + \frac{1}{4} (\underline{u}_1 + \Delta t \partial_t \underline{u}_1), \\ \underline{u}^+ &= \frac{1}{3} \underline{u} + \frac{2}{3} (\underline{u}_2 + \Delta t \partial_t \underline{u}_2). \end{aligned} \quad (154)$$

10.1 Well-balancedness and entropy conservation

In this test case, the well-balancedness and entropy conservation properties of the two-parameter family of fluxes (49) and corresponding volume (85) and surface terms (92), where the parameters are chosen according to (100) and the free parameters $m_4, k_9, k_{10}, k_{11}, l_{10}$ have been set to zero, are investigated.

The domain $[-1, 1]$ is equipped with periodic boundary conditions and the solution is evolved in the time interval $[0, 1]$ using 1000 steps of the SSPRK(3,3) (154) method. The bottom topography is given by

$$b(x) = \sin \frac{\pi x}{4}, \quad (155)$$

and the initial condition is chosen as

$$h_0(x) = 1 - b(x), \quad hv_0(x) = 0 \quad (156)$$

for the lake-at-rest test case. Using $N = \frac{120}{p+1} = 15$ elements with polynomials of degree $\leq p = 7$, represented using Gauß nodes, the simulations have been performed for $(a_1, a_2) \in \left\{ -3 + \frac{k}{10}, 0 \leq k \leq 60 \right\}^2$ with gravitational constant $g = 1$.

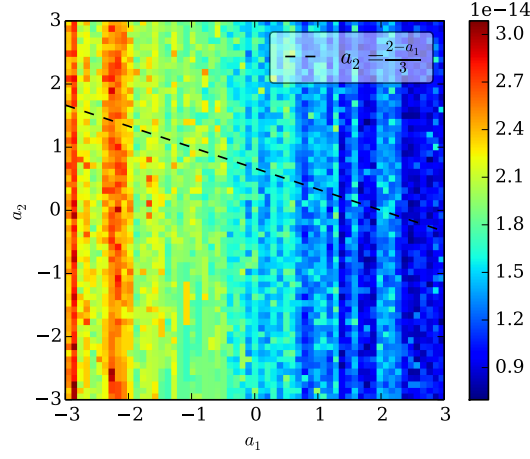


Figure 2: Maximum norm error $\max \left\{ \|h(1) - h_0\|_\infty, \|hv(1) - hv_0\|_\infty \right\}$ of solutions computed using the entropy conservative fluxes with varying parameters a_1, a_2 for the lake-at-rest initial condition (156).

The results, visualised in Figure 2 show the excellent well-balancedness of the methods. The maximum errors $\max \left\{ \|h(1) - h_0\|_\infty, \|hv(1) - hv_0\|_\infty \right\}$ (computed at the nodes) are of order 10^{-14} using Float64 (i.e. double precision) in Julia 0.4.7 (Bezanson, Edelman, Karpinski, and Shah, 2014).

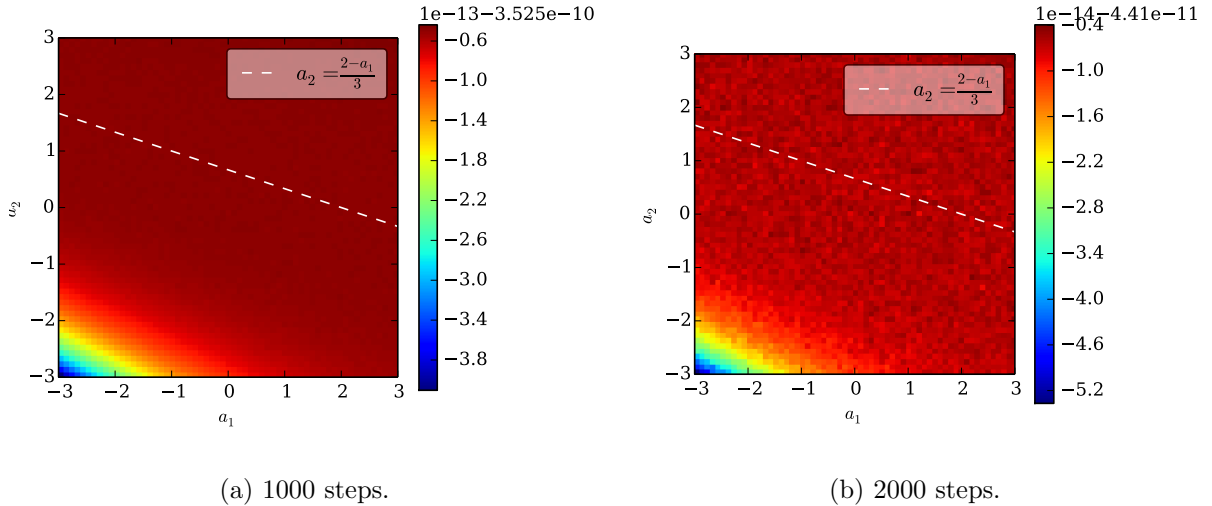


Figure 3: Relative entropy dissipation $(\int U(1) - \int U(0)) / \int U(0)$ of solutions computed using the entropy conservative fluxes with varying parameters a_1, a_2 for the initial condition (157).

Choosing the initial condition

$$h_0(x) = 1, \quad hv_0(x) = 0, \quad (157)$$

$h + b$ is no longer constant, but the solution remains smooth at first. Computing again until $t = 1$ with the same parameters as before, the loss of entropy is visualised in Figure 3. Using 1000 steps of SSPRK(3,3) (154), the relative entropy dissipation $(\int U(1) - \int U(0)) / \int U(0)$ is of order 10^{-10} , with variations of order 10^{-13} for different parameters a_1, a_2 . This loss of entropy is caused by the time integrator, as can be seen by refining the time step. Using 2000 steps, the relative dissipation is order order 10^{-11} with variations of order 10^{-14} . The smooth solutions for $a_1 = -1, a_2 = \frac{2-a_1}{3} = 1$ are plotted in Figure 4.

The influence of the numerical (surface) flux is visualised in Figure 5. There, the number of degrees of freedom $N \cdot (p + 1)$ has been kept constant, while the polynomial degree p varies between 0 (first order FV scheme) and 5. As can be seen there, the entropy conservative flux $f^{-1,1}$

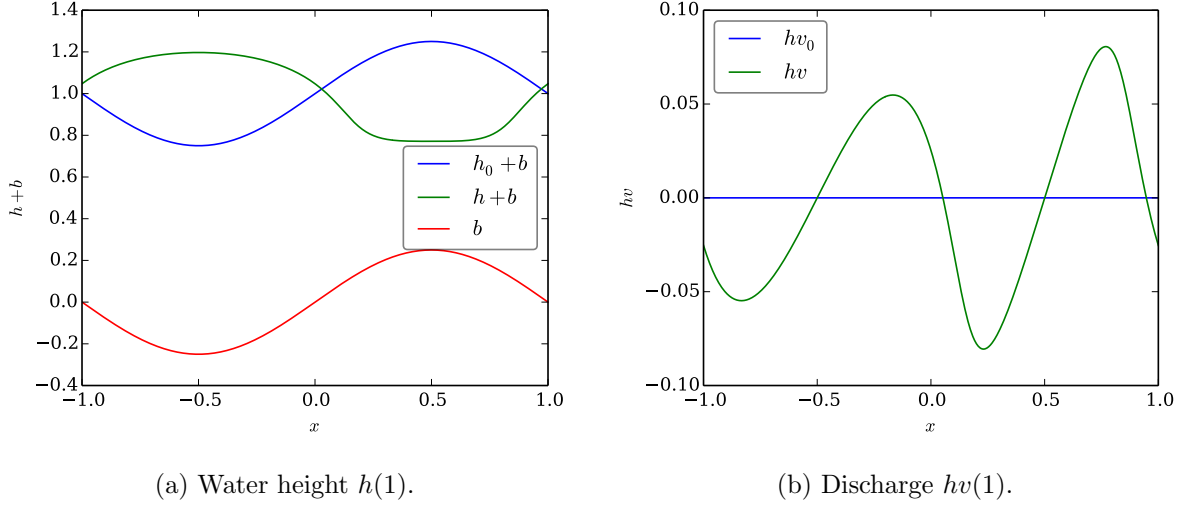


Figure 4: Solutions computed using the entropy conservative fluxes with parameters $a_1 = -1, a_2 = \frac{2-a_1}{3} = 1$ for the initial condition (157).

is indeed entropy conservative, while the entropy stable fluxes are a bit dissipative. The dissipation increases from the Suliciu flux (141) over the kinetic flux (144) to the local Lax-Friedrichs flux (134). All three fluxes have been implemented using the hydrostatic reconstruction of Audusse, Bouchut, Bristeau, Klein, and Perthame (2004) described in section 8.6.

In the finite volume setting $p = 0$, the dissipation is of order 10^{-3} and decreases with increasing polynomial degree p . For $p \geq 3$, the curves for the dissipative fluxes become visually indistinguishable and for $p \geq 5$ they coincide with the entropy conservative flux $f^{-1,1}$ for this smooth solution.

10.2 Lake at rest with emerged bump

Here, the lake-at-rest initial condition of SWASHES (Delestre, Lucas, Ksinant, Darboux, Laguerre, Vo, James, Cordier, et al., 2013, Section 3.1.2)

$$b(x) = \begin{cases} 0.2 - 0.05(x - 10)^2, & \text{if } 8 < x < 12, \\ 0, & \text{else,} \end{cases} \quad (158)$$

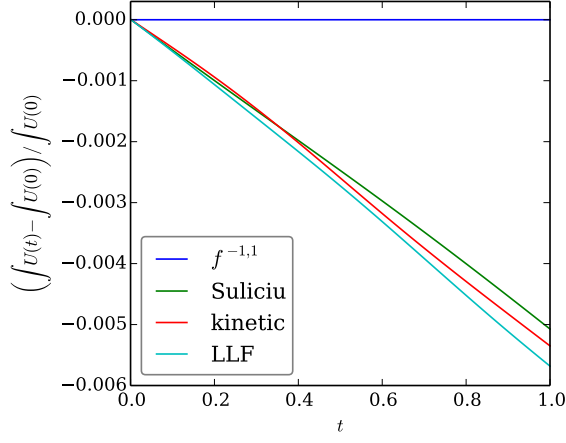
$$h_0(x) = \max \{0.1, b(x)\} - b(x), \quad hv_0(x) = 0,$$

will be used in the domain $[0, 25]$ with periodic boundary conditions for simulations in the time interval $[0, 1]$ with gravitational constant $g = 9.81$.

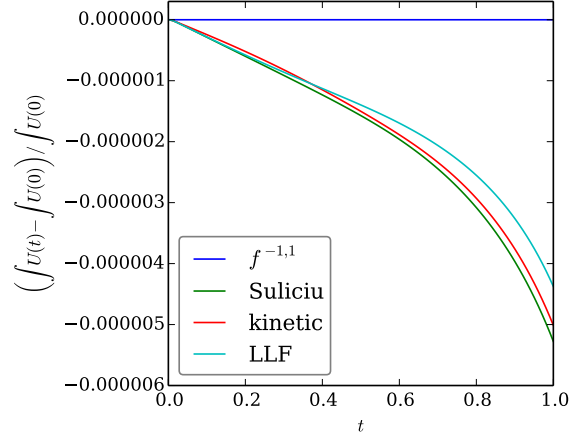
If no FV subcells are used, the result strongly depends on the resolution of the shore and needs in general some additional dissipation to be stable near the wet-dry front. However, activating FV subcells if the water height h at some node in the element is smaller than 10^{-5} , the simulation is stable.

These results are shown in Figure 6 for $N = 40$ elements of polynomials of degree $\leq p = 5$ and the local Lax-Friedrichs flux (134) with hydrostatic reconstruction as numerical flux.

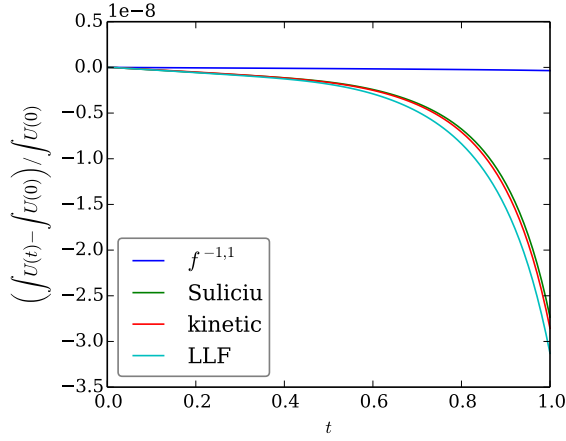
The maximum error norm $\max \{ \|h(1) - h_0\|_\infty, \|hv(1) - hv_0\|_\infty \}$ (computed at the nodes) is of order of magnitude 10^{-16} for varying parameters $(a_1, a_2) \in \left\{ -3 + \frac{k}{10}, 0 \leq k \leq 60 \right\}^2$ used for the volume terms (85). Again, Gauß nodes and corresponding surface terms have been used, where the additional free parameters have been set to zero. Additionally, the water height for the choice $a_1 = -1, a_2 = \frac{2-a_1}{3} = 1$ is visualised there.



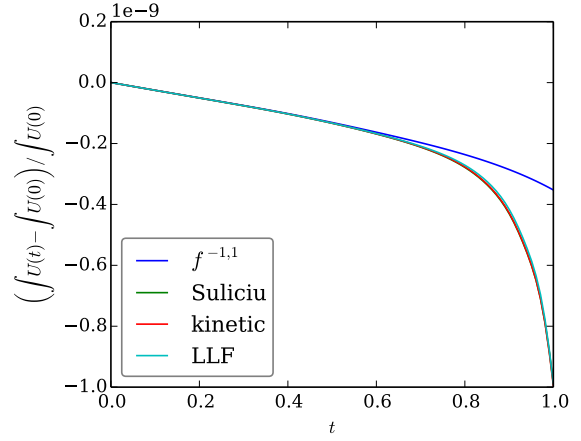
(a) $p = 0$, $N = \frac{120}{p+1} = 120$.



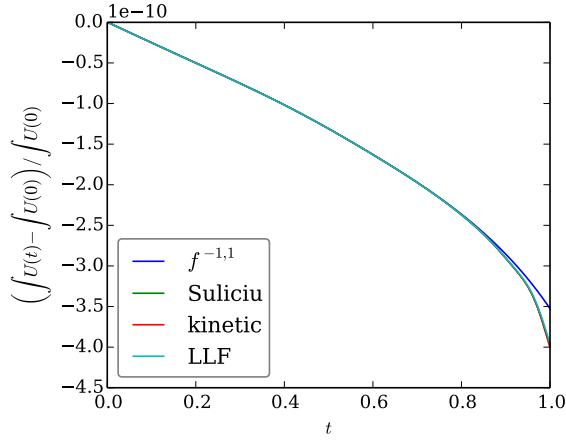
(b) $p = 1$, $N = \frac{120}{p+1} = 60$.



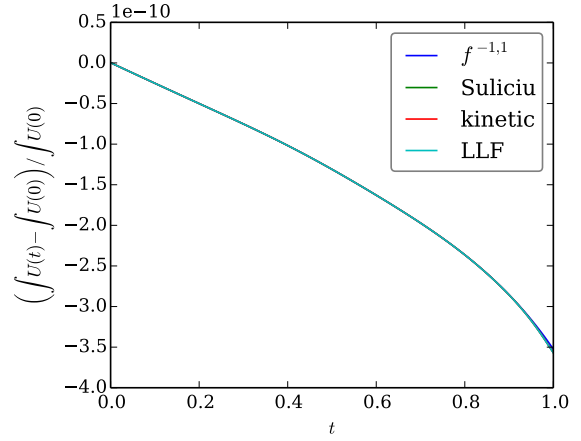
(c) $p = 2$, $N = \frac{120}{p+1} = 40$.



(d) $p = 3$, $N = \frac{120}{p+1} = 30$.



(e) $p = 4$, $N = \frac{120}{p+1} = 20$.



(f) $p = 5$, $N = \frac{120}{p+1} = 15$.

Figure 5: Relative entropy dissipation $(\int U(1) - \int U(0)) / \int U(0)$ of solutions computed using different surface fluxes for the initial condition (157) with varying degree p and number of elements $N = \frac{120}{p+1}$.

10.3 Moving water equilibrium with varying bottom b

Here, a moving water equilibrium of the shallow water equations with gravitational constant $g = 9.81$ given by

$$hv \equiv m = \text{const}, \quad \frac{1}{2}v^2 + g(h + b) \equiv E = \text{const} \quad (159)$$

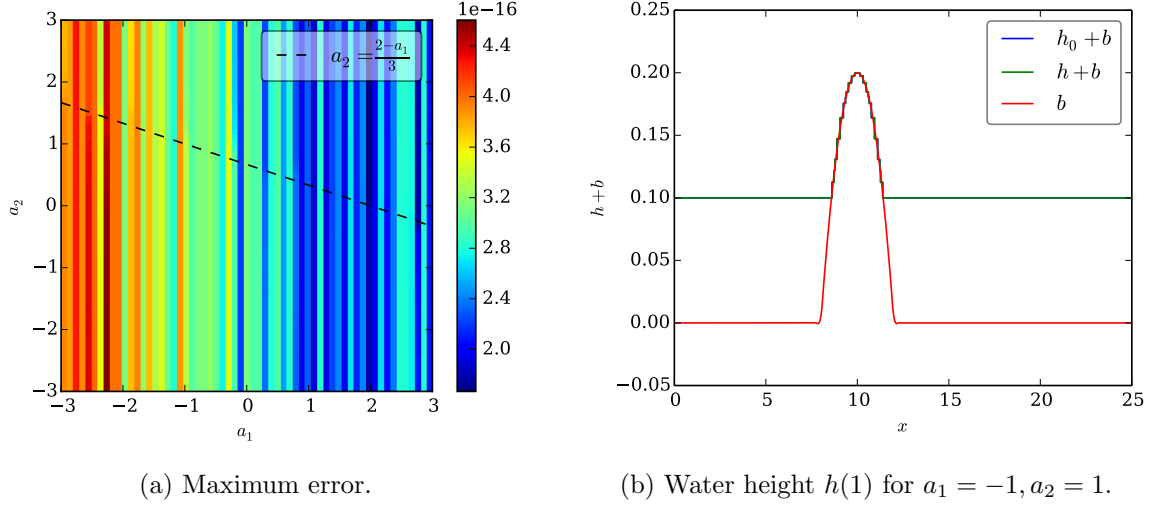


Figure 6: Maximum norm error $\max \left\{ \|h(1) - h_0\|_\infty, \|hv(1) - hv_0\|_\infty \right\}$ of solutions computed using the entropy conservative fluxes with varying parameters a_1, a_2 for the lake-at-rest with emerged bump initial condition (158) and solution at $t = 1$ for $a_1 = -1, a_2 = 1$.

is considered. The bottom topography is

$$b(x) = \begin{cases} \frac{1}{4} \cos(10\pi(x+1)) + \frac{1}{4}, & \text{if } -0.1 < x < 0.1, \\ 0, & \text{else,} \end{cases} \quad (160)$$

and the initial condition is computed by solving the second equation of (159) for h , inserting $v^2 = \frac{(hv)^2}{h^2} = \frac{m^2}{h^2}$. Two initial conditions $m = 1, E = 25$ and $m = 3, E = \frac{3}{2}(mg)^{2/3} + \frac{g}{2} = 19.203311922761937$ are considered, similar to Audusse, Chalons, and Ung (2015).

Computing the maximum error $\max \left\{ \|h(1) - h_0\|_\infty, \|hv(1) - hv_0\|_\infty \right\}$ at the nodes yields identical results for both initial conditions with polynomial degrees $\leq p \in \{0, \dots, 9\}$, and parameters $(a_1, a_2) \in \left\{ -3 + \frac{k}{10}, 0 \leq k \leq 60 \right\}^2$ for the volume terms, while the local Lax-Friedrichs flux (134) has been used as numerical flux. The domain is divided into $N = 40$ elements using Gauß nodes.

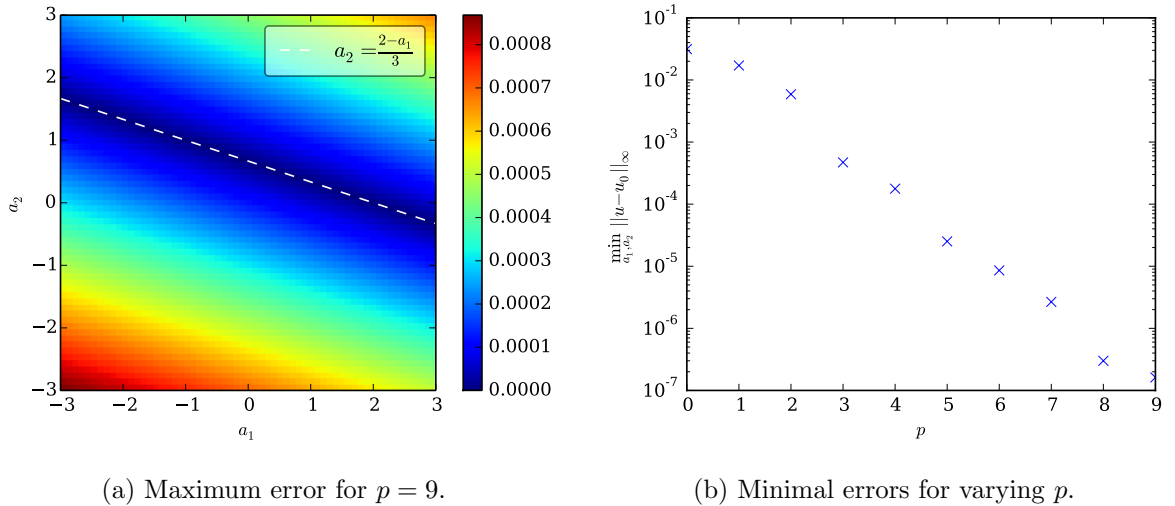


Figure 7: Maximum norm error $\max \left\{ \|h(1) - h_0\|_\infty, \|hv(1) - hv_0\|_\infty \right\}$ of solutions computed using the entropy conservative fluxes with varying parameters a_1, a_2 for moving water equilibrium (159) with $m = 1, E = 25$ for polynomials of degree $\leq p = 9$ and minimal values of the maximum errors over a_1, a_2 .

These results with $m = 1, E = 25$ are shown in Figure 7. As can be seen there, the choice

$a_2 = \frac{2-a_1}{3}$ is optimal for this problem, while the choice of a_1 does not seem to be critical. This can be explained by the additional term in v for $a_2 \neq \frac{2-a_1}{3}$ in the numerical flux (49) and the corresponding volume terms (85). The results for $m = 3, E = \frac{3}{2}(mg)^{2/3}$ are visually indistinguishable.

Additionally, the minimal values of the maximum error over the parameters a_1, a_2 are plotted in Figure 7 for $m = 1, E = 25$. The usual superior properties of odd polynomial degrees p as well as exponential convergence can be seen there.

10.4 Dam break

Here, the dam break problem with dry domain of SWASHES (Delestre, Lucas, Ksinant, Darboux, Laguerre, Vo, James, Cordier, et al., 2013, Section 4.1.2) will be considered. The initial condition

$$h_0(x) = \begin{cases} h_l = 0.005, & \text{if } x < 5, \\ 0, & \text{else,} \end{cases} \quad hv_0(x) = 0, \quad b(x) = 0, \quad (161)$$

is evolved in the domain $[0, 10]$ until $t = 6$, and the gravitational constant is again $g = 9.81$. The analytical solution is given by

$$h(t, x) = \begin{cases} h_l, & \text{if } x < 5 - \sqrt{gh_l}, \\ \frac{4}{9g} \left(\sqrt{gh_l} - \frac{x-5}{2t} \right)^2, & \text{if } 5 \leq x < 5 + 2t\sqrt{gh_l}, \\ 0, & \text{else,} \end{cases} \quad (162)$$

$$v(t, x) = \begin{cases} 0, & \text{if } x < 5 - \sqrt{gh_l}, \\ \frac{2}{3} \left(\frac{x-5}{t} + \sqrt{gh_l} \right), & \text{if } 5 \leq x < 5 + 2t\sqrt{gh_l}, \\ 0, & \text{else,} \end{cases}$$

where again $h_l = 0.005$. The results of a simulation using $N = 100$ elements with polynomials of degree $\leq p = 2$ and the local Lax-Friedrichs numerical flux (134) are plotted in Figure 8. Here, FV subcells are used in a cell if the water height in the cell itself or adjacent cells is less than 10^{-6} , and the parameters are chosen as $a_1 = -1, a_2 = \frac{2-a_1}{3} = 1$.

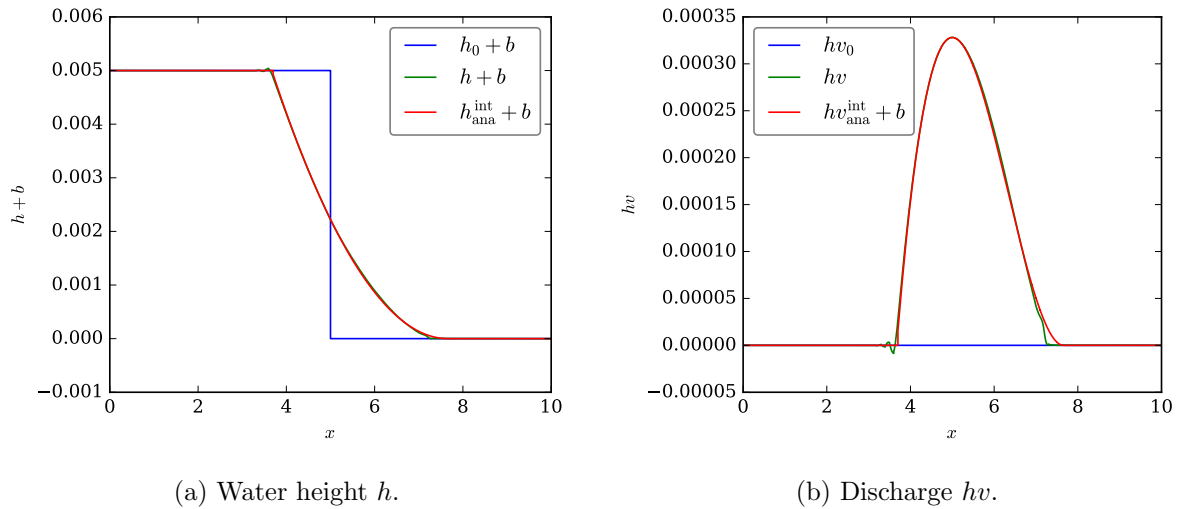


Figure 8: Numerical solution computed using $N = 100$ elements of polynomials of degree $\leq p = 2$ on Gauß nodes for the dam break problem (161) with local Lax-Friedrichs numerical flux and $a_1 = -1, a_2 = 1$.

Motivated by the result of section 10.3, only the parameter a_1 has been varied for this problem, while the parameter a_2 is fixed at $a_2 = \frac{2-a_1}{3}$. the results for $a_1 \in \left\{ -3 + \frac{k}{10}, 0 \leq k \leq 60 \right\}$ are shown in Figures 9 and 10. There, the L_2 errors have been computed exactly for the polynomials using Gauß nodes and the $\|\cdot\|_\infty$ errors are computed at the same nodes.

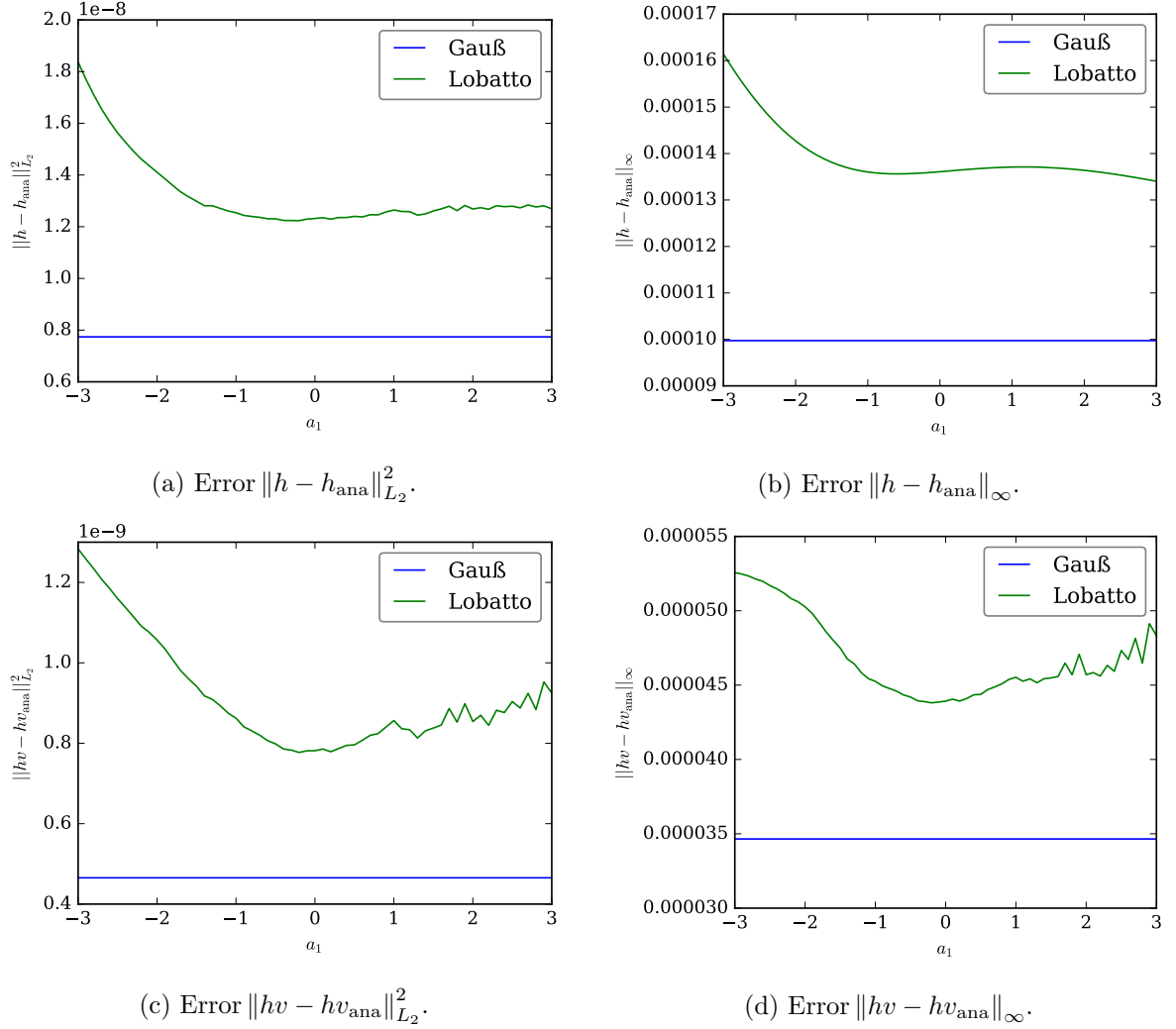


Figure 9: Errors of the numerical solution for varying parameter a_1 and $N = 50$ elements of polynomials of degree $\leq p = 2$ using Gauß and Lobatto nodes.

In these experiments, Gauß nodes yield a lower error in the solutions, both in $\|\cdot\|_{L_2}$ and $\|\cdot\|_{\infty}$ and this error is nearly independent of the parameter a_1 (it varies at most three orders of magnitude lower). However, the error using Lobatto nodes are influenced by the choice of a_1 with variations up to 50%.

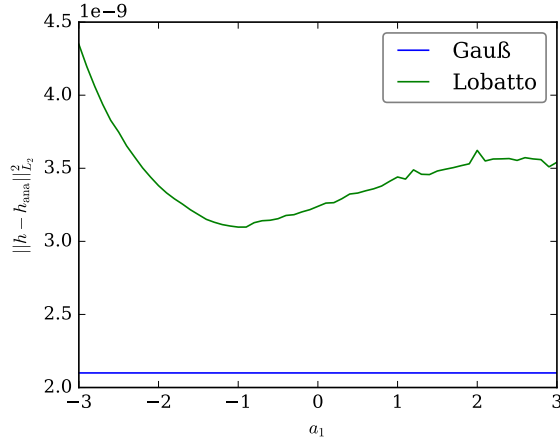
The corresponding errors in both norms $\|\cdot\|_{L_2}$ and $\|\cdot\|_{\infty}$ follow approximately the same trend for h and hv , respectively, but there are differences between the error curves of the height h and the discharge hv .

These results, especially the ones for the discharge hv , suggest, that choosing the parameter a_1 between -1 and 0 might be optimal, but this has to be investigated thoroughly.

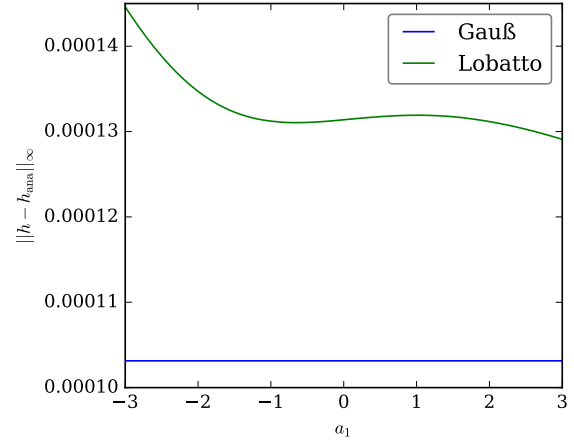
11 Summary and conclusions

A two-parameter family of entropy stable and well-balanced numerical fluxes and corresponding split forms with adapted surface terms for general SBP bases including Lobatto and Gauß nodes has been developed. The positivity preserving framework of Zhang and Shu (2011) can be used in this setting, but has to be accompanied by some additional dissipation / stabilisation mechanism near wet-dry fronts. Here, the subcell finite volume framework has been used and extended naturally to diagonal-norm nodal SBP bases.

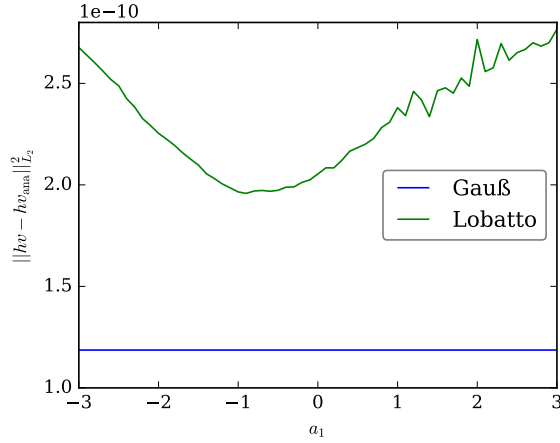
Numerical tests confirm the properties of the derived schemes. As suggested by a first physicists intuition, the second parameter of the two-parameter family should be chosen as $a_1 = \frac{2-a_1}{3}$ in order not to use some higher order terms in the velocity v . This choice has been advantageous for the considered moving water equilibrium in section 10.3.



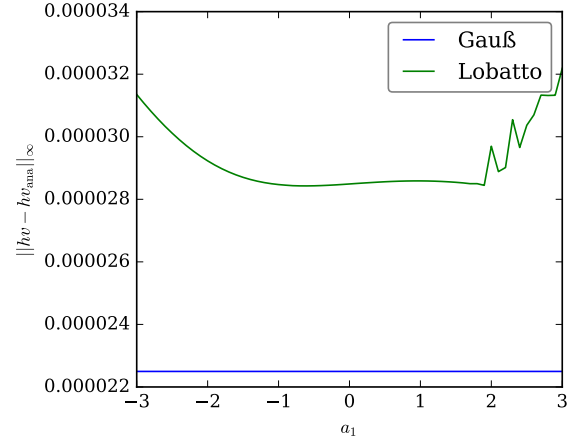
(a) Error $\|h - h_{\text{ana}}\|_{L_2}^2$.



(b) Error $\|h - h_{\text{ana}}\|_{\infty}$.



(c) Error $\|hv - hv_{\text{ana}}\|_{L_2}^2$.



(d) Error $\|hv - hv_{\text{ana}}\|_{\infty}$.

Figure 10: Errors of the numerical solution for varying parameter a_1 and $N = 100$ elements of polynomials of degree $\leq p = 2$ using Gauß and Lobatto nodes.

However, the choice of the first parameter a_1 does not seem to be similarly simple. There is no clear physical intuition at first and the dam break experiments in section 10.4 are not unambiguous. Thus, further analytical and numerical studies have to be performed in order to understand the influence of this parameter and possible optimal choices.

Additional topics of further research include the extension to curvilinear coordinates in several space dimensions similarly to Wintermeyer, Winters, Gassner, and Kopriva (2016) and the investigation of interactions of curved elements with the parameter a_1 , of other means performing finite volume subcell projection, and other stabilisation techniques.

References

- Audusse, E., F. Bouchut, M.-O. Bristeau, R. Klein, and B. Perthame (2004). “A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows”. In: *SIAM Journal on Scientific Computing* 25.6, pp. 2050–2065.
- Audusse, E., C. Chalons, and P. Ung (2015). “A simple well-balanced and positive numerical scheme for the shallow-water system”. In: *Commun. Math. Sci* 13.5, pp. 1317–1332.
- Barth, T. J. (1999). “Numerical methods for gasdynamic systems on unstructured meshes”. In: *An introduction to recent developments in theory and numerics for conservation laws*. Springer, pp. 195–285.
- Berthon, C. and C. Chalons (2016). “A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations”. In: *Mathematics of Computation* 85.299, pp. 1281–1307.

- Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah (2014). *Julia: A fresh approach to numerical computing*. arXiv:1411.1607 [cs.MS].
- Bouchut, F. (2003). “Entropy satisfying flux vector splittings and kinetic BGK models”. In: *Numerische Mathematik* 94.4, pp. 623–672.
- Bouchut, F. (2004). *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Springer Science & Business Media.
- Delestre, O., S. Cordier, F. Darboux, and F. James (2012). “A limitation of the hydrostatic reconstruction technique for Shallow Water equations”. In: *Comptes Rendus Mathématique* 350.13, pp. 677–681.
- Delestre, O., C. Lucas, P.-A. Ksinant, F. Darboux, C. Laguerre, T.-N. Vo, F. James, S. Cordier, et al. (2013). “SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies”. In: *International Journal for Numerical Methods in Fluids* 72.3. (used corrected version from arXiv), pp. 269–300. arXiv:1110.0288v7 [math.NA].
- Duran, A. and F. Marche (2014). “Recent advances on the discontinuous Galerkin method for shallow water equations with topography source terms”. In: *Computers & Fluids* 101, pp. 88–104.
- Einfeldt, B. (1988). “On Godunov-type methods for gas dynamics”. In: *SIAM Journal on Numerical Analysis* 25.2, pp. 294–318.
- Einfeldt, B., C.-D. Munz, P. L. Roe, and B. Sjögren (1991). “On Godunov-type methods near low densities”. In: *Journal of Computational Physics* 92.2, pp. 273–295.
- Fernández, D. C. D. R., J. E. Hicken, and D. W. Zingg (2014). “Review of summation-by-parts operators with simultaneous approximation terms for the numerical solution of partial differential equations”. In: *Computers & Fluids* 95, pp. 171–196.
- Fisher, T. C. and M. H. Carpenter (Feb. 2013). *High-Order Entropy Stable Finite Difference Schemes for Nonlinear Conservation Laws: Finite Domains*. Technical Report NASA/TM-2013-217971. NASA Langley Research Center, Hampton VA 23681-2199, United States: NASA.
- Fisher, T. C., M. H. Carpenter, J. Nordström, N. K. Yamaleev, and C. Swanson (2013). “Discretely conservative finite-difference formulations for nonlinear conservation laws in split form: Theory and boundary conditions”. In: *Journal of Computational Physics* 234, pp. 353–375.
- Fjordholm, U. S., S. Mishra, and E. Tadmor (2011). “Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography”. In: *Journal of Computational Physics* 230.14, pp. 5587–5609.
- Frid, H. (2001). “Maps of Convex Sets and Invariant Regions for Finite Difference Systems of Conservation Laws”. In: *Archive for Rational Mechanics and Analysis* 160.3, pp. 245–269.
- Frid, H. (2004). “Correction to “Maps of Convex Sets and Invariant Regions for Finite Difference Systems of Conservation Laws””. In: *Archive for Rational Mechanics and Analysis* 171.2, pp. 297–299.
- Gassner, G. J. (2013). “A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to SBP-SAT finite difference methods”. In: *SIAM Journal on Scientific Computing* 35.3, A1233–A1253.
- Gassner, G. J., A. R. Winters, and D. A. Kopriva (2016a). “A well balanced and entropy conservative discontinuous Galerkin spectral element method for the shallow water equations”. In: *Applied Mathematics and Computation* 272, pp. 291–308.
- Gassner, G. J., A. R. Winters, and D. A. Kopriva (2016b). *Split Form Nodal Discontinuous Galerkin Schemes with Summation-By-Parts Property for the Compressible Euler Equations*. arXiv:1604.06618 [math.NA].
- Gottlieb, S. and C.-W. Shu (1998). “Total variation diminishing Runge-Kutta schemes”. In: *Mathematics of Computation* 67.221, pp. 73–85.
- Harten, A., P. D. Lax, and B. van Leer (1983). “On upstream differencing and Godunov-type schemes for hyperbolic conservation laws”. In: *SIAM Review* 25.1, pp. 35–61.
- Holden, H. and N. H. Risebro (2002). *Front tracking for hyperbolic conservation laws*. Vol. 152. Springer.

- Liu, X.-D. and S. Osher (1996). “Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes I”. In: *SIAM Journal on Numerical Analysis* 33.2, pp. 760–779.
- Meister, A. and S. Ortleb (2016). “A positivity preserving and well-balanced DG scheme using finite volume subcells in almost dry regions”. In: *Applied Mathematics and Computation* 272, pp. 259–273.
- Ortleb, S. (2016). *Kinetic energy preserving DG schemes based on summation-by-parts operators on interior node distributions*. Talk presented at the joint annual meeting of DMV and GAMM. TU Braunschweig.
- Perthame, B. and C. Simeoni (2001). “A kinetic scheme for the Saint-Venant system with a source term”. In: *Calcolo* 38.4, pp. 201–231.
- Ranocha, H., P. Öffner, and T. Sonar (2015). *Extended skew-symmetric form for summation-by-parts operators*. Submitted. arXiv:1511.08408 [math.NA].
- Ranocha, H., P. Öffner, and T. Sonar (2016). “Summation-by-parts operators for correction procedure via reconstruction”. In: *Journal of Computational Physics* 311, pp. 299–328. arXiv:1511.02052 [math.NA].
- Svärd, M. and J. Nordström (2014). “Review of summation-by-parts schemes for initial-boundary-value problems”. In: *Journal of Computational Physics* 268, pp. 17–38.
- SymPy Development Team (2016). *SymPy: Python library for symbolic mathematics*. Version 1.0. URL: <http://www.sympy.org>.
- Tadmor, E. (1987). “The numerical viscosity of entropy stable schemes for systems of conservation laws. I”. In: *Mathematics of Computation* 49.179, pp. 91–103.
- Tadmor, E. (2003). “Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems”. In: *Acta Numerica* 12, pp. 451–512.
- Wintermeyer, N., A. R. Winters, G. J. Gassner, and D. A. Kopriva (2016). *An Entropy Stable Nodal Discontinuous Galerkin Method for the Two Dimensional Shallow Water Equations on Unstructured Curvilinear Meshes with Discontinuous Bathymetry*. arXiv:1509.07096v2 [math.NA].
- Wolfram Research, Inc. (2014). *Mathematica*. Version 10.0. URL: <https://www.wolfram.com>.
- Xing, Y. and C.-W. Shu (2014). “A survey of high order schemes for the shallow water equations”. In: *Journal of Mathematical Study* 47.221-249, p. 56.
- Xing, Y., X. Zhang, and C.-W. Shu (2010). “Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations”. In: *Advances in Water Resources* 33.12, pp. 1476–1493.
- Zhang, X. and C.-W. Shu (2011). “Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments”. In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 467. 2134. The Royal Society, pp. 2752–2776.